МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет *компьютерных наук*
Кафедра *теории обработки и защиты информации*

Алгоритмы стегоанализа изображений с использованием глубоких нейронных сетей и их программная реализация

ВКР *Магистерская диссертация*
09.04.02 Информационные системы и технологии
Безопасность информационных систем

Допущено к защите в ГЭК

Зав. кафедрой	_____	*А. А. Сирота, д. т. н., проф.* __.__.20__
Обучающийся	_____	*Н. А. Нагорный, 2 курс, д/о*
Руководитель	_____	*А. А. Сирота, д. т. н., проф.*

Воронеж 2019

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Факультет компьютерных наук

Кафедра теории обработки и защиты информации

УТВЕРЖДАЮ
заведующий кафедрой

_____ _____
подпись, расшифровка подписи
__.__.2019

ЗАДАНИЕ
НА ВЫПОЛНЕНИЕ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ
НАГОРНОГО НИКИТЫ АЛЕКСЕЕВИЧА

1. Тема работы «Алгоритмы стегоанализа изображений с использованием глубоких нейронных сетей и их программная реализация» утверждена решением ученого совета факультета компьютерных наук от __ .__.2019
2. Направление подготовки / специальность 09.04.02 Информационные системы и технологии
3. Срок сдачи законченной работы __.__ 2019
4. Календарный план: (строится в соответствии со структурой ВКР)

№	Структура ВКР	Сроки выполнения	Примечание
	Введение	**01.10.2017— 03.10.2017**	
	Глава 1. *Анализ предметной области.Обзор известных работ в области классического стегоанализа*	**01.12.2017— 01.03.2018**	
	Глава 2. *Разработка и реализация алгоритмов стегоанализа изображений с использованием свёрточных искусственных нейронных сетей*	**03.03.2018— 01.06.2018**	
	Глава 3. *Описание разработанного программного обеспечения и экспериментальных исследований*	**01.03.2019— 01.06.2019**	
	Заключение	**02.06.2019**	
	Список используемых источников	**03.06.2019**	

Обучающийся _____ _____
Подпись *расшифровка подписи*

Руководитель _____ _____
Подпись *расшифровка подписи*

Реферат

Магистерская диссертация 55 с., 21 рис.

СТЕГОАНАЛИЗ, СТЕГАНОГРАФИЯ, МАШИННОЕ ОБУЧЕНИЕ, СВЁРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ

Объектом исследования являются алгоритмы стегоанализа изображений в оттенках серого с использованием свёрточных нейронных сетей.

Цель работы — разработка алгоритма стегоанализа цветных изображений с использованием свёрточных нейронных сетей и его программная реализация.

В ходе выполнения работы был проведён анализ существующих подходов к стегоанализу изображений в оттенках серого и архитектур свёрточных нейронных сетей, а также осуществлена разработка программной реализации рассмотренных архитектур.

В результате исследования предложен собственный алгоритм стегоанализа цветных изображений и разработана его программная реализация.

Содержание

Введение

На протяжении последнего десятилетия стегоанализ изображений является одним из активно развивающихся направлений в области цифровой стеганографии. Его главной задачей является обнаружение присутствия стегосообщений в цифровых изображениях. Существует множество разнообразных стегоаналитических подходов, но наибольшей эффективностью характеризуется группа методов, основанных на статистической обработке анализируемых данных. Их целью является построение математической модели заполненного контейнера, однако эта задача крайне осложнена разнообразием стеганографических методов и вносимых ними искажающих воздействий на контейнеры. В качестве такой модели, в основном, используется многомерное множество признаков, а решение задачи стегоанализа сводится к выполнению двух шагов: извлечению существенных признаков стегоконтейнеров, позволяющих судить о наличии стегосигнала, и бинарной классификации, сопоставляющей входные объекты с классами «пустой стегоконтейнер» и «заполненный стегоконтейнер». Успех стегоанализа в такой форме напрямую зависит от того, насколько точно извлечённые признаки отражают искажающий эффект стеговстраивания.

Активный рост вычислительных мощностей открывает новые возможности для создания адаптивных алгоритмов стеговстраивания, стремящихся обеспечить наилучшие характеристики стегосистемы применительно к конкретному стегоконтейнеру. Это влечёт за собой усложнение статистических зависимостей между отдельными элементами изображения, что, в свою очередь, существенно усложняет задачу создания математической модели заполненного стегоконтейнера и ручное конструирование признаков.

Математический аппарат глубоких искусственных нейронных сетей способен решить эту проблему путём автоматизации процесса извлечения признаков. Важным преимуществом таких признаков является их высокая

релевантность, обеспечиваемая наличием в искусственной нейронной сети обратной связи между классификатором и экстрактором. К тому же, использование нейронных сетей значительно ускоряет процесс конструирования признаков и обеспечивает стегоаналитика набором неочевидных характерных особенностей заполненных стегоконтейнеров.

Актуальной является задача построения стегодетектора с использованием свёрточных нейронных сетей, способного работать с цветными изображениями, поэтому разработка соответствующего стегоаналитического алгоритма и стала целью данной работы.

Для достижения поставленной цели необходимо решить следующие задачи:

– проанализировать предметную область и изучить существующие стегоаналитические подходы;

– разработать программную реализацию существующих алгоритмов стегоанализа, использующих свёрточные нейронные сети;

– разработать собственный алгоритм стегоанализа цветных изображений на основе рассмотренных алгоритмов стегоанализа изображений в оттенках серого;

– реализовать один или несколько методов стеговстраивания для тестирования алгоритма;

– провести эксперимент и сравнить разработанный алгоритм с существующими.

1 Анализ предметной области. Обзор известных работ в области классического стегоанализа

1.1 Базовые понятия и положения стеганографии

Стеганография (от греч. στεγανός — скрытый и γράφω — пишу; буквально «тайнопись») — это наука о передаче (хранении) информации с сохранением в тайне самого факта передачи (хранения). Большинство способов классической цифровой стеганографии базируется на особенностях восприятия информации человеком, организуя сокрытие секретных сообщений таким образом, что чувствительность человеческих органов не позволяет определить их наличие.

Основными понятиями стеганографии являются понятия контейнера, сообщения и ключа. Ниже приведены их определения, данные в монографии [1].

Контейнером b ($b \in \mathbb{B}$, где $\mathbb{B}$ — множество всех возможных контейнеров) называют несекретные данные, используемые для сокрытия сообщений. В цифровой стеганографии роль контейнеров играют растровые графические изображения, цифровой звук, а также текстовые и другие электронные документы.

Сообщением m ($m \in \mathbb{M}$, где $\mathbb{M}$ — множество всех возможных сообщений) называют секретную информацию, скрываемую в контейнере.

Ключом k ($k \in \mathbb{K}$, где $\mathbb{K}$ — множество всех возможных секретных ключей) называют секретную информацию, известную только санкционированному пользователю стеганосистемы, которая определяет конкретный способ сокрытия и извлечения сообщения. В широком смысле ключ — это неизвестный противнику способ (алгоритм) сокрытия информации, в узком — параметр заранее оговорённого стеганографического алгоритма, без знания которого извлечение сообщения невозможно.

Пустой контейнер — это некоторый контейнер b, не содержащий какого-либо сообщения m.

Заполненный контейнер — это контейнер b, содержащий сообщение m.

Помимо приведённых определений распространены следующие термины.

Стеганографическая система (стегосистема) — совокупность средств, осуществляющих встраивание сообщения в контейнер и его последующее извлечение.

Пропускная способность стегосистемы C — отношение максимально возможного объёма встроенного сообщения к объёму контейнера при сохранении необходимого уровня стегоустойчивости.

Стегоустойчивость — свойство стегосистемы, характеризующее её способность обеспечивать заданное качество стегосвязи в условиях наличия различных деструктивных факторов (шумы, атаки противодействия передаче сообщения).

Стегоскрытность — свойство стегосистемы, характеризующее её способность обеспечивать невозможность факта наличия стегосвязи.

Общая схема стеганографической системы приведена на рисунке 1.1. Согласно ей, на передающей стороне сообщение скрывается в контейнере при помощи прямого стеганографического преобразования. Затем полученный заполненный контейнер по открытому каналу связи отправляется принимающей стороне, где при помощи обратного стеганографического преобразования извлекается исходное сообщение.

Задача *пассивного противника* состоит в определении факта наличия в контейнере сокрытых данных (атака детектирования). *Активный противник* пытается вносить изменения в контейнер или, в простейшем случае, уничтожить передаваемое сообщение (атака уничтожения данных).

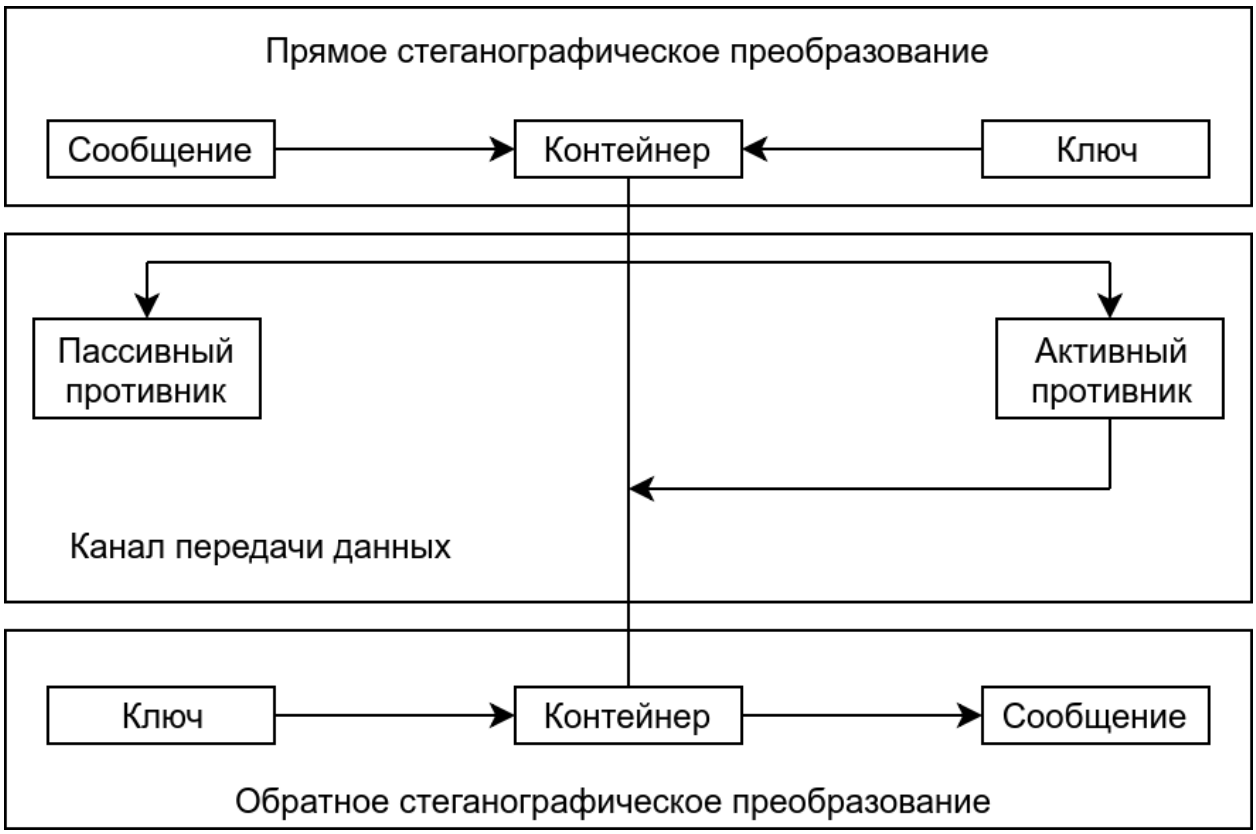


Рисунок 1.1 – Обобщённая схема стеганографической системы

1.2 Метод замены наименее значимого бита

Наиболее распространённым и примитивным методом сокрытия данных в пространственной области изображений является метод замены наименее значимого бита (англ. Least Significant Bit replacing; также метод замены НЗБ). Одним из его основных преимуществ является низкая вычислительная сложность, основным недостатком — низкая устойчивость к атакам детектирования и уничтожения данных.

Обозначим пустой стегоконтейнер в оттенках серого размера $n_1 \times n_2$ как $X = (x_{ij}) \in \mathbb{Q} = \{0, \ldots, 255\}^{n_1 \times n_2}$, а соответствующий заполненный контейнер как $Y = (y_{ij}) \in \mathbb{Q}$.

Младший бит пикселя изображения в оттенках серого хранит младший разряд соответствующего двоичного числа и, следовательно, его изменение минимально влияет на изменение числа в целом. Отсюда следует, что встраивание произвольной битовой последовательности в младшие биты пикселей

повлечёт за собой минимальной визуальное искажение стегоконтейнера. К тому же, такая методика встраивания обеспечивает довольно высокую пропускную способность $C = 1$ бит/пиксель.

Пустой контейнер можно представить в двоичной форме следующим образом:

$$(x_{ij}) = \sum_{p=0}^{8} (x_{ij})_p \cdot 2^p.$$

Тогда заполненный стегоконтейнер имеет вид:

$$(y_{ij}) = \sum_{p=1}^{8} (x_{ij})_p \cdot 2^p + m_{ij},$$

где $m_{ij} \in 0, 1$ — бит стегосообщения, имеющего вид матрицы размера $n_1 \times n_2$.

Существует множество нетривиальных модификаций данного метода. Интересным для рассмотрения примером является алгоритм, описанный в [2]. Он предназначен для стеговстраивания цифровых водяных знаков. В таком случае контейнер является не только средой передачи сообщения, которую можно выбрать для обеспечения необходимого уровня скрытности и пропускной способности относительно конкретного сообщения, но и несёт собственную ценность, что повышает требования к уровню искажений, вносимых при стеговстраивании. К тому же, в некоторых областях применения (например, в медицинской отрасли) наличие искажений, приемлемых для передачи сообщений, недопустимо. Рассматриваемый метод учитывает это и обеспечивает обратимость внесения искажений.

Первым шагом операции встраивания является разделение пикселей (x_{ij}) на непересекающиеся группы из n пикселей $G = (x_1, x_2, \ldots, x_n)$. Для примера выбираются группы из $n = 4$ последовательных пикселей в строке. Затем определяется различающая функция $f(x_1, x_2, \ldots, x_n) \in \mathbb{R}$, имеющая

следующий вид:

$$f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n-1} |x_{i+1} - x_i|.$$

Назначением данной функции является оценка гладкости или «регулярности» группы пикселей G.

Следующим этапом является применение обратимой операции сдвига F, модифицирующей яркость пикселя на близкое значение. Например, операцию F, меняющую НЗБ, можно определить следующим образом: $0 \leftrightarrow 1, 2 \leftrightarrow 3, \ldots, 254 \leftrightarrow 255$. Обозначим данную операцию F_1. Тогда F_{-1} применяется по следующей схеме: $-1 \leftrightarrow 0, 1 \leftrightarrow 2, 3 \leftrightarrow 4, \ldots, 253 \leftrightarrow 254, 255 \leftrightarrow 256$. Для полноты введём $F_0(x) = x \quad \forall x \in \mathbb{Q}$.

Различающая функция f и операция сдвига F применяются для классификации групп пикселей G:

$$\begin{cases} G \in \mathbb{R} \iff f(F(G)) > f(G), \\ G \in \mathbb{S} \iff f(F(G)) < f(G), \\ G \in \mathbb{U} \iff f(F(G)) = f(G). \end{cases} \tag{1.1}$$

Группы, принадлежащие ко множествам $\mathbb{R}$, $\mathbb{S}$ и $\mathbb{U}$ называются регулярными, сингулярными и неиспользуемыми соответственно. N_R, N_S и N_U — количество упомянутых групп в изображении.

Из 1.1 видно, что:

$$\begin{cases} \forall G \in \mathbb{R} \quad F(G) \in S, \\ \forall G \in \mathbb{S} \quad F(G) \in R, \\ \forall G \in \mathbb{U} \quad F(G) \in U. \end{cases}$$

Таким образом, возможно ввести соответствие между $G \in \mathbb{R}$ и единичным битом встраиваемого сообщения и $G \in \mathbb{S}$ и нулевым битом, применяя F в случае несоответствия типа группы очередному биту сообщения. Для восстановления оригинального изображения предлагается составить карту

принадлежности всех G изображения множествам и разместить её в начале встраиваемого сообщения.

Пропускная способность данного метода $C = N_R + N_S = n_1 n_2 / n - N_U$. Количество неиспользуемых групп напрямую влияет на пропускную способность, что существенно ограничивает длину сообщения при встраивании в астрономические снимки и другие изображения с продолжительными участками пикселей равной яркости. Для устранения этого недостатка предлагается использовать различные операторы сдвига F для пикселей группы G, а для записи соответствия между оператором и пикселем — маску M, представляющую собой последовательность из n чисел, принимающих значения -1, 0 и 1. Тогда $F_M(G) = (F_{M(1)}(x_1), F_{M(2)}(x_2), \ldots, F_{M(n)}(x_n))$.

1.3 Основные принципы и подходы стегоанализа

Стегоанализ — это наука о выявлении факта передачи скрытой информации в анализируемых данных, её извлечении или уничтожении.

В зависимости от наличия априорных знаний о стегосистеме, задействованной при стеговстраивании, применяются направленные или слепые методы стегоанализа. *Направленные* методы стегоанализа предназначены для выявления присутствия сообщения, вложенного известным стегоаналитику методом. В случае отсутствия данных о стегоалгоритме встраивания используют методы *слепого* стегоанализа.

Следующий критерий классификации стегоаналитических методов — используемый подход. В зависимости от подхода можно выделить *визуальные*, *сигнатурные*, *схемные* и *статистические* методы. Рассмотрим их применительно к стегоанализу цифровых неподвижных изображений.

Визуальные методы основаны на способностях зрительной системы человека, а именно, возможности анализировать зрительные образы и определять различия в сопоставляемых изображениях.

Такие методы стегоанализа графических файлов являются самыми простыми в реализации, но они неспособны обнаружить сообщения, встроенные с помощью современных методов стеговстраивания, и подвержены влиянию человеческого фактора в лице оператора-аналитика.

Подходом **сигнатурных методов** является синтаксический анализ последовательности битов анализируемого контейнера.

В первых известных алгоритмах сокрытия сообщений в цифровых изображениях заполнение контейнера производилось путём замены служебных атрибутов файла на встраиваемую информацию [3], что позволяло относительно просто составить сигнатуру, характеризующую заполненный стегоконтейнер, и свести задачу стегоанализа к задаче поиска сигнатуры.

Дальнейшее усовершенствование стеганографических методов привело к широкому распространению стегосистем, осуществляющих встраивание в пространственной области изображения, но сигнатурные методы всё ещё могли противостоять таким стегосистемам. Например, в работе [4] описывается сигнатурный метод стегоанализа изображений в оттенках серого, стеганографически заполненных с помощью программы Hide and Seek версий 4.1 и 5.0. Заполненный контейнер представляет собой изображение, содержащее три канала цветности (цветовая модель RGB) с одинаковыми значениями яркости в каждом. Наиболее распространённым способом хранения трёхканальных изображений является кодирование каждой цветовой координаты пикселя 8-битным числом, определяющим диапазон уровней квантования $\mathbb{Q} = \{0, \ldots, 255\}$ и шаг квантования $h = 1$. Hide and Seek использует $\mathbb{Q} = \{0, \ldots, 252\}$ и $h = 4$, что позволяет выделить контейнеры, заполненные этой программой, на общем фоне благодаря меньшему количеству градаций яркости и отсутствию абсолютно белого цвета. Столь грубый подход также приводит к высокой визуальной заметности стеганографического искажения, вплоть до очевидного повреждения изображения (рисунок 1.2).

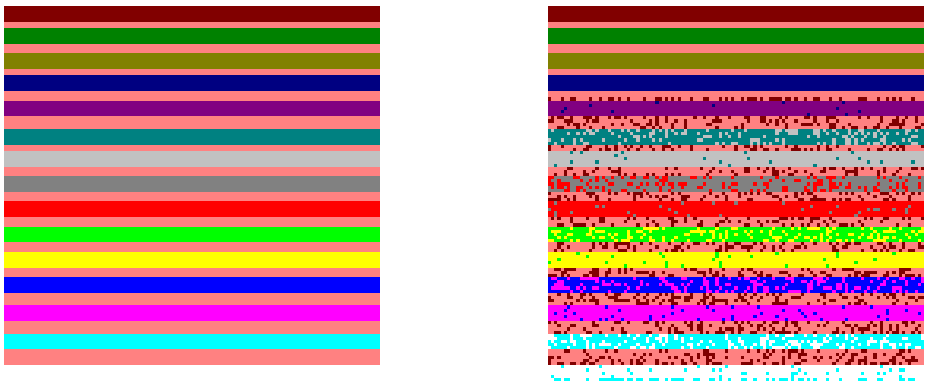


Рисунок 1.2 – Пустой и заполненный в Hide and Seek контейнеры

Преимуществом данного подхода является высокая скорость работы реализующих его алгоритмов, но сильная привязка к атрибутам заполненных стегоконтейнеров, не учитывающая принципы стеговстраивания, является существенным недостатком, не позволяющим широко использовать данный подход.

Схемные методы применяются для проверки гипотез о наличии стеганографического встраивания с применением известной стегосистемы.

Основными достоинствами схемных методов являются низкая вероятность возникновения ошибок и возможность идентификации стеганографической системы с последующим извлечением сообщения.

Фатальными недостатками данного подхода являются невозможность реализовать необходимое для применения слепого стегоанализа количество стегосистем и многообразие ключей, позволяющее стегосистеме менять характер стеговстраивания от одного ключа к другому.

Статистические методы базируются на понятии «естественного» контейнера. Их суть заключается в оценивании вероятности существования стегосообщения, встроенного неизвестной стегосистемой на основе критерия близости исследуемого контейнера к «естественному» путём обнаружении отклонения анализируемой информации от ожидаемой модели.

Преимуществом подхода является возможность применения в слепом стегоанализе, недостатком — высокая сложность создания модели «естественного» контейнера.

1.4 Известные методы стегоанализа на основе статистической обработки данных

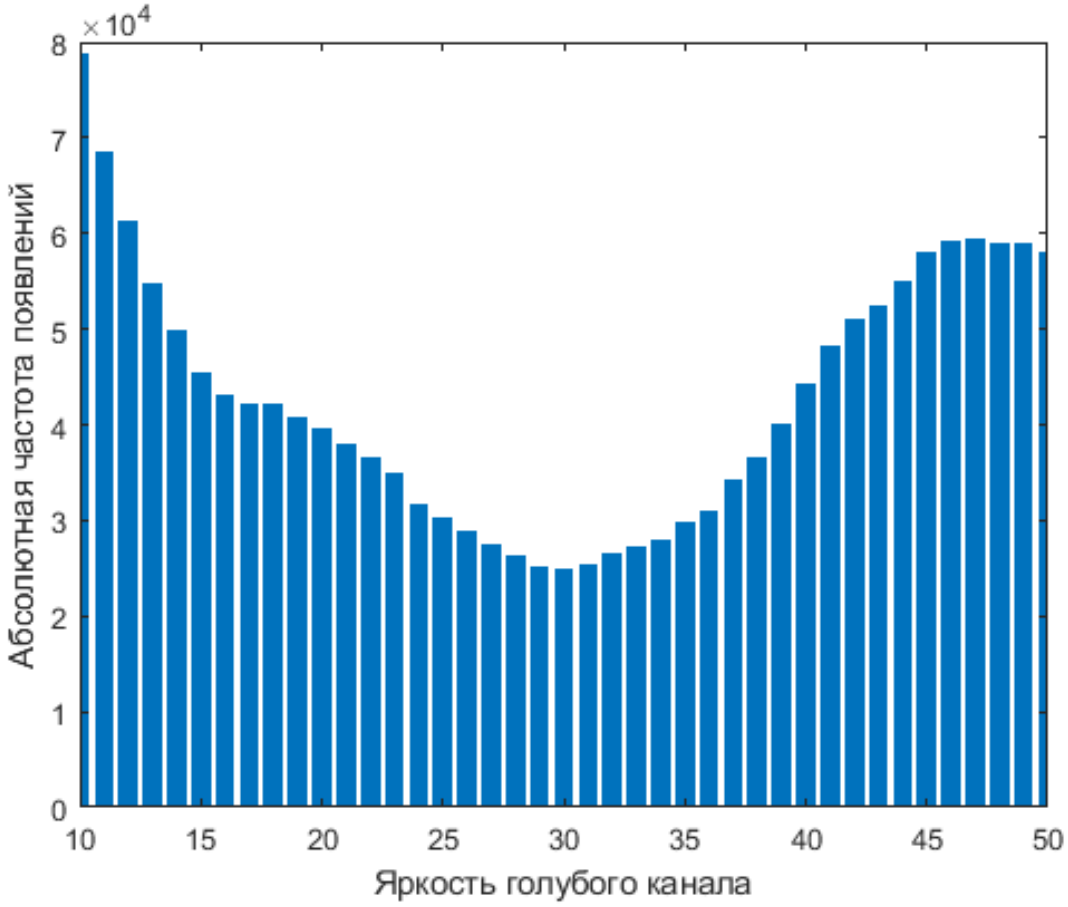
1.4.1 Метод анализа гистограмм

Метод анализа гистограмм применяется при наличии гипотезы о распределении элементов пустого контейнера (x_{ij}). Для анализируемого контейнера (b_{ij}), о заполненности которого нет априорного знания, строят такую же гистограмму. По отклонению распределения, полученного для (b_{ij}), от распределения, характерного для (x_{ij}), судят о заполненности (b_{ij}). Примером атаки на стегосистему, использующую метод замены наименее значимого бита, может служить построение гистограммы яркости пикселей. Рисунок 1.3 наглядно иллюстрирует эффект группировки различных значений яркости, вызванный заменой НЗБ и выражающейся в ступенчатом характере гистограммы.

Данный метод также используется для анализа частоты появления серий k наименее значимых битов значений яркости пикселей. Для построения гистограммы подсчитывается частота равенства последнего НЗБ нулю и единице ($k = 1$), двух НЗБ сериям-двойкам (00, 01, 10, 11; $k = 2$), трёх НЗБ сериям-тройкам (000, 001, 010, 011, 100, 101, 110, 111; $k = 3$) и т. д. Для заполненных контейнеров характерна близость частот, в пустых данная симметрия не наблюдается.

1.4.2 RS-анализ

RS-анализ был предложен в [5] и основывается на модификации метода замены НЗБ, описанного в разделе 1.2.



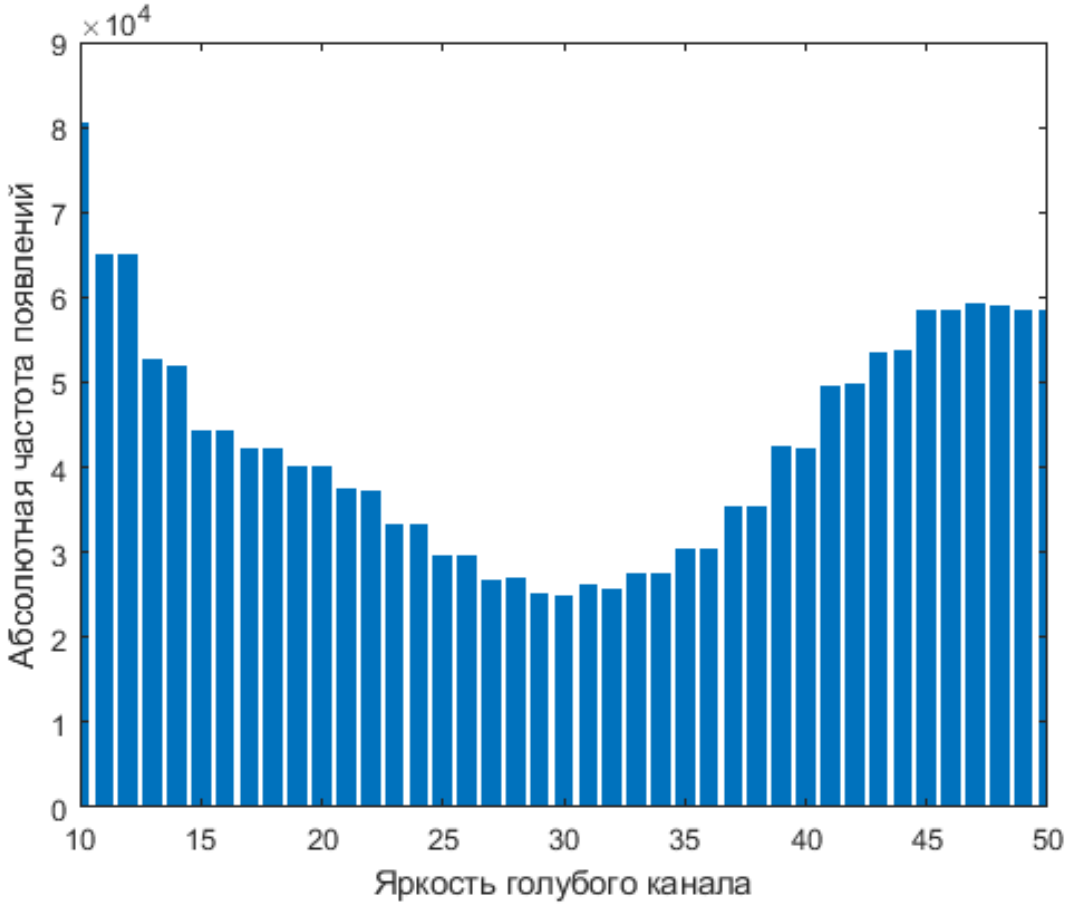


Рисунок 1.3 – Гистограммы до и после применения метода замены НЗБ

Введём обозначение R_M для относительного количества регулярных групп с маской M среди всех групп изображения. Аналогично, обозначим относительное количество сингулярных групп S_M. Тогда $R_M + S_M \leq 1$ и $R_{-M} + S_{-M} \leq 1$ для обратной маски. Статистическая гипотеза данного стегоаналитического метода утверждает, что для пустого контейнера

$$\begin{cases} R_M \approx R_{-M}, \\ S_M \approx S_{-M}. \end{cases} \tag{1.2}$$

Экспериментально выяснено, что для изображений с рандомизированными НЗБ пикселей равенства 1.2 не соблюдаются. Это наглядно иллюстрирует рисунок 1.4.

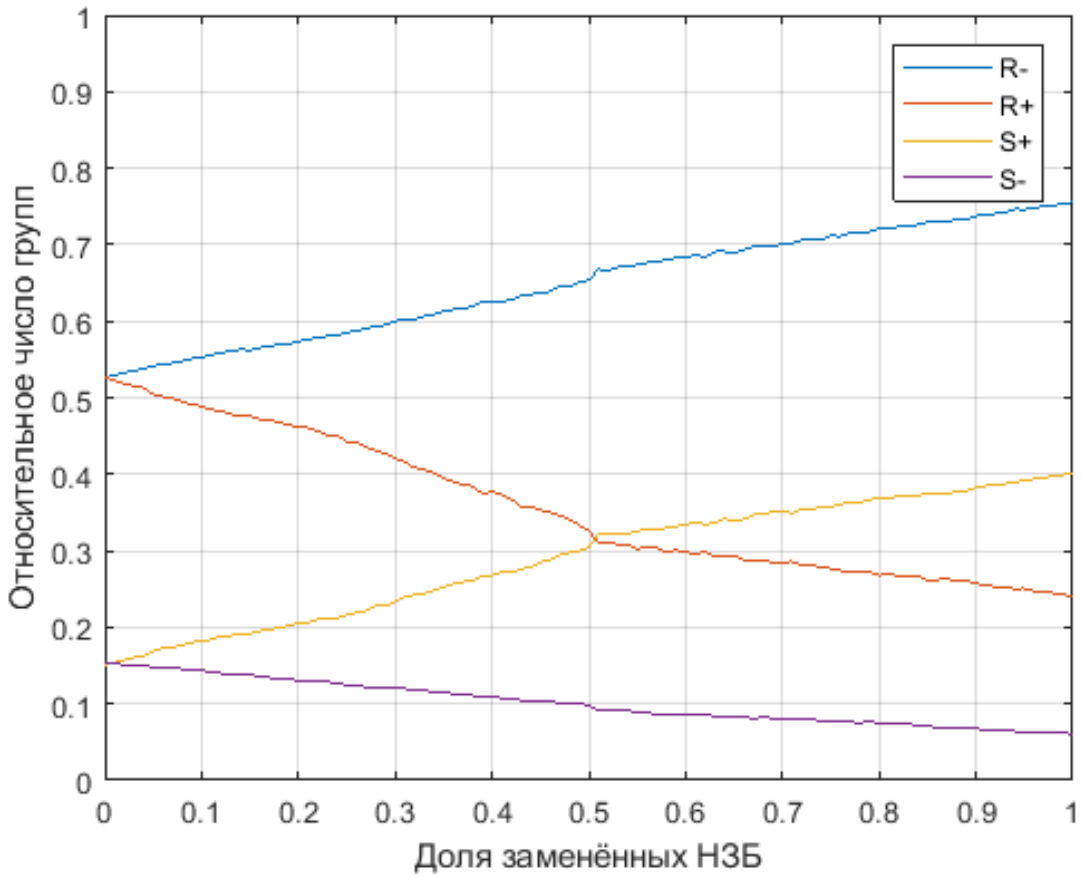


Рисунок 1.4 – Пример RS-диаграммы

RS-диаграмма представляет собой график зависимости относительного числа групп различного типа от доли заменённых НЗБ. Диаграмма,

изображённая на рисунке 1.4 показывает, что пары R_M — R_{-M} и S_M — S_{-M} действительно очень близки для пустых контейнеров. Также видно, что разница между R_M и S_M стремится к нулю при увеличении длины p сообщения m, пока оно не заполнит 50 % пикселей. Для R_{-M} и S_{-M} можно наблюдать обратный эффект: разница между ними растёт с увеличением p.

Принцип данного стегоаналитического метода заключается в построении RS-диаграммы и поиске точек пересечения пар кривых путём решения уравнения

$$2(d_1 + d_0)x^2 + (d_{-0} - d_{-1} - d_1 - 3d_0)x + d_0 - d_{-0} = 0,$$

где

$$\begin{cases} d_0 = R_M(p/2), -S_M(p/2), \\ d_{-0} = R_{-M}(p/2), -S_{-M}(p/2), \\ d_1 = R_M(1 - p/2), -S_M(1 - p/2), \\ d_{-1} = R_{-M}(1 - p/2), -S_{-M}(1 - p/2). \end{cases}$$

Корень уравнения x позволяет найти p:

$$p = x/(x - 1/2).$$

1.5 Машина опорных векторов в задачах стегоанализа

С появлением более совершенных стегосистем методы статистического стегоанализа уже не обеспечивали достаточной точности детектирования стегосообщений в контейнерах. Нарушение статистических зависимостей между элементами исходных контейнеров стало настолько незначительным, что судить о заполненности или пустоте контейнера, опираясь на вид гистограммы или небольшое количество статистических критериев не представлялось возможным. Возникла необходимость агрегировать результаты статистической обработки. Решением данной задачи стало использование машинного обучения и машин опорных векторов, в частности.

Машина опорных векторов (МОВ; англ. SVM, support vector machine) — набор алгоритмов обучения с учителем, использующихся для классификации. Различают три основных типа МОВ:

а) линейная сеперабельная МОВ,

б) линейная несеперабельная МОВ,

в) нелинейная МОВ.

Рассмотрим общий случай процедуры обучения с учителем: на вход обучаемой системы подаётся набор тренировочных примеров, который обычно называют *обучающим* или *тренировочным набором данных*, состоящих из пар «*стимул — реакция*». Задача обучаемой системы состоит в том, чтобы дополнить стимулы из *тестового набора данных*, не участвовавшего в обучении, соответствующими реакциями, руководствуясь взаимосвязями между стимулами и реакциями обучающего набора. Работоспособность подхода обеспечивается подобностью данных, доступных для обучения, и данных, на которых впоследствии применяется обученная система.

Задачи обучения с учителем обычно делятся на задачи *классификации* и *регрессии*. В задаче классификации поданный на вход объект соотносится с одним из классов, в задаче регрессии требуется предсказать значение некоей функции, у которой обычно может быть бесконечно много разных решений.

В задачах классификации роль стимула играет объект, а роль реакции — класс объекта. В стегоаналитических приложениях в качестве объектов обычно выступают контейнеры (исходные или предварительно обработанные) либо *признаки* — заранее выбранные характеристики контейнеров, совокупность которых позволяет судить об их принадлежности к классу пустых либо классу заполненных контейнеров.

Пусть $\boldsymbol{x} \in X$ — вектор признаков контейнера, а $y \in Y$ — номер класса. $\forall X, Y \quad \exists y^* : X \to Y$. Значения $y^* : X \to Y$ априорно известны для контейнеров тренировочной выборки $X^\ell = (\boldsymbol{x}_i, y_i), i = 1, \ldots, \ell; y_i = y^*(\boldsymbol{x}_i)$.

Алгоритм $a : X \to Y$ аппроксимирует целевую зависимость y^* на всём пространстве X.

Для задачи классификации на два непересекающихся класса, в котором объекты описываются n-мерными вещественными векторами, $X = \mathbb{R}^n$, $Y = \{-1, +1\}$. Алгоритм линейного порогового классификатора имеет вид:

$$a(\boldsymbol{x}) = sign(\sum_{j=1}^{n} w_j x^j - w_0) = sign(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - w_0),$$

где $\boldsymbol{x} = (x^1, \ldots, x^n) \in \mathbb{R}^n$. Вектор $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ и скалярный порог $w_0 \in \mathbb{R}$ являются параметрами алгоритма.

Уравнение $\langle \boldsymbol{w}, \boldsymbol{x} \rangle = w_0$ описывает гиперплоскость размерности $n - 1$, разделяющую классы в пространстве $\mathbb{R}^n$.

Предположим, что выборка разделима, то есть существуют такие значения параметров w, w_0, при которых функционал числа ошибок

$$Q(w, w_0) = \sum_{i=1}^{l} [y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0) < 0]$$

принимает нулевое значение. Тогда разделяющая гиперплоскость не единственна: существуют другие её положения, реализующие то же самое разбиение выборки. Для достижения наименьшего значения $Q(w, w_0)$ МОВ накладывает на разделяющую гиперплоскость требование соблюдения максимальной дистанции от ближайших к ней точек обоих классов. Таким образом, задаётся полоса, разделяющая классы, в которой не лежат точки обучающей выборки. Границами полосы служат две параллельные гиперплоскости, на которых лежат точки, ближайшие к разделяющей гиперплоскости. При этом сама разделяющая гиперплоскость проходит ровно по середине полосы.

В общем случае гарантировать линейную разделимость выборки не представляется возможным, поэтому на практике используют линейный несеперабельный МОВ, позволяющий алгоритму допускать ошибки и стремящийся к минимизации их количества.

Существует ещё один подход к решению проблемы линейной неразделимости — переход от исходного пространства признаковых описаний объектов X к новому пространству H с помощью некоторого преобразования $\psi : X \to H$. Если пространство имеет достаточно высокую размерность, то можно надеяться, что в нём выборка окажется линейно разделимой (если выборка X^l не противоречива, то всегда найдётся пространство размерности не более l, в котором выборка будет линейно разделима). Пространство H называют *спрямляющим*. Пример результата обучения нелинейной МОВ приведён на рисунке 1.5.

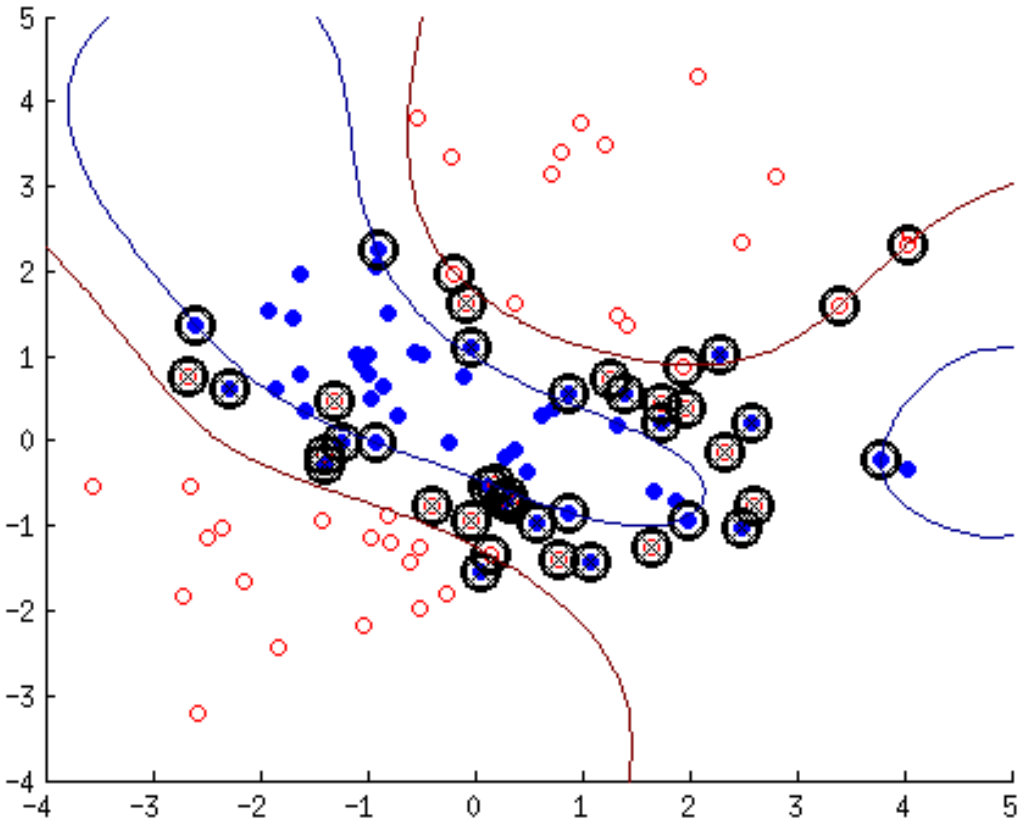


Рисунок 1.5 – Результат обучения нелинейной МОВ

Анализ контейнера при использовании данного метода производится в несколько этапов:

а) выбор характеристик стегообъектов, позволяющих судить о пустоте или заполненности контейнера и называемых *признаками*;

б) обучение машины опорных векторов на достаточно большой выборке пустых и заполненных стегоконтейнеров;

в) применение обученной машины для классификации.

Конструирование пространства признаков является частным случаем задачи построения математической модели заполненного стегоконтейнера, наличие которой позволило бы легко вычислять принадлежность анализируемых объектов к тому или иному классу. Зачастую в качестве таких признаков используются статистические характеристики.

Пример, иллюстрирующий применение МОВ в задачах стегоанализа описан в работе [6]. Авторы предлагают стегоаналитический алгоритм SHDFT:

а) цветной контейнер b разделяется на каналы b_R, b_G и b_B;

б) для каждого из них строится гистограмма частот появления значений яркости, для которой находятся математическое ожидание и дисперсия (всего 6 признаков);

в) полученные гистограммы подвергается дискретному преобразованию Фурье, для результата которого рассчитываются математическое ожидание, дисперсия, коэффициенты асимметрии и эксцесса, а также энергия сигнала во всех трёх каналах (всего 15 признаков);

г) рассчитывается математическое ожидание разности гистограммы частот яркостей и результата дискретного преобразования Фурье (всего 3 признака);

д) пункты 1-4 повторяются для всей обучающей выборки, используемой для обучения МОВ.

Подход, демонстрирующий использование признаков из областей преобразования исходного изображения, продемонстрирован в [7] производится кадрирование набора изображений на фрагменты размером 16×16 пикселей. Каждый фрагмент подвергается дискретному косинусному и вейвлет-

преобразованиям, таким образом, в трёх областях рассчитывается математическое ожидание, дисперсия, коэффициенты асимметрии и эксцесса.

2 Разработка и реализация алгоритмов стегоанализа изображений с использованием свёрточных искусственных нейронных сетей

Машина опорных векторов дала мощный толчок появлению всё более совершенных стегоаналитических методов, однако со временем стегоаналитики ощутили несовершенство практики ручного выделения признаков, и для автоматизации данной процедуры стали использоваться свёрточные нейронные сети.

Пространства признаков, построенные нейронными сетями, по многим параметрам превосходят сконструированные вручную. Благодаря автоматизации их извлечения, размерность пространства признаков зависит не от возможностей стегоаналитика, а от вычислительной мощности используемого аппаратного обеспечения. К тому же, многие признаки довольно сложно получить аналитически ввиду неочевидности статистических зависимостей между элементами контейнеров. Свёрточные нейронные сети на примере задач распознавания образов доказывают свою способность к построению таких признаков. Наконец, метод обратного распространения ошибки [8] реализует обратную связь, позволяющую конструировать признаки непосредственно во время обработки тренирующей выборки с учётом ошибки классификации, что исключает временные затраты на итеративную процедуру формирования пространства признаков целиком, его тестирования и последующего исправления.

2.1 Описание нейросетевого подхода и общей схемы алгоритма

Одной из первых работ, положивших начало нейросетевому подходу к решению задач классификации, стал линейный перцептрон Фрэнка Розенблатта [9, 10]. По сути своей перцептрон Розенблатта — это линейная модель бинарной классификации, задача которой — научиться сопоставлять метки классов не участвовавшим в обучении объектам.

Будем считать, что каждый вход представляет собой вектор вещественных чисел $\boldsymbol{x} = (x^1, x^2, \ldots, x^n) \in \mathbb{R}^n$, и входы в тренировочном множестве снабжены известными выходами $y(\boldsymbol{x}) \in \{-1, +1\}$. Тогда для решения задачи необходимо найти такие веса $w_0, w_1, \ldots, w_n \in \mathbb{R}^n$, чтобы знак линейной функции

$$sign(w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_n x^n) \tag{2.1}$$

как можно чаще совпадал с правильным ответом $y(\boldsymbol{x})$. Положительное значение 2.1 интерпретируют как суждение о принадлежности $\boldsymbol{x}$ к классу с меткой $+1$, и наоборот.

Для удобства увеличим размерность вектора $\boldsymbol{x}$ таким образом, чтобы он принял вид $\boldsymbol{x} = (1, x^1, x^2, \ldots, x^n) \in \mathbb{R}^{n+1}$. Это позволит считать линейную комбинацию $w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_n x^n$ скалярным произведением $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$, где $\boldsymbol{w} = (w_0, w_1, w_2, \ldots, w_n)$. Архитектура однослойного перцептрона представлена на рисунке 2.1.

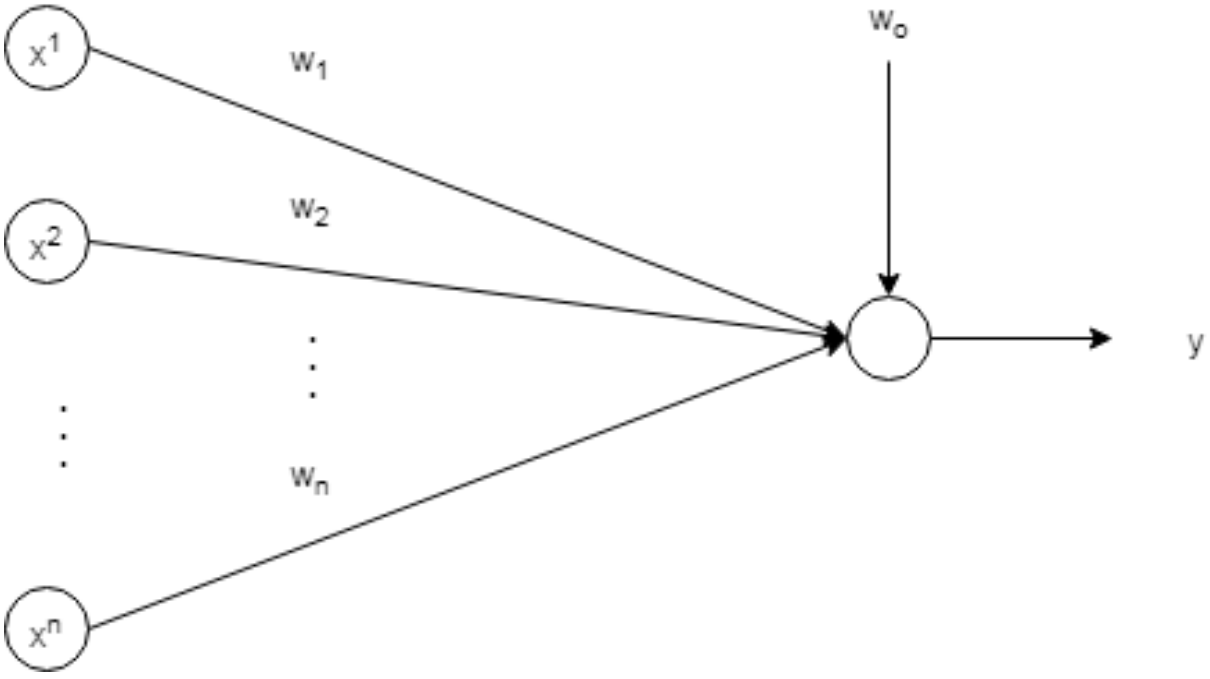


Рисунок 2.1 – Архитектура однослойного перцептрона

Для обучения данной функции необходимо выбрать функцию ошибки. Розенблатт вводит в этом качестве *критерий перцептрона*:

$$E_P(\boldsymbol{w}) = -\sum_{\boldsymbol{x}\in\mathcal{M}} y(\boldsymbol{x})\langle \boldsymbol{w}, \boldsymbol{x}\rangle, \tag{2.2}$$

где $\mathcal{M}$ обозначает множество множество примеров, которые перцептрон с весами **w** классифицирует неверно.

Оптимизировать 2.2 можно методом градиентного спуска, согласно которому вектор весов на шаге обучения τ имеет следующий вид:

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta\nabla_{\boldsymbol{w}} E_P(\boldsymbol{w}).$$

Такой классификатор способен работать только с линейно разделимыми множествами. Популярным антипримером использования перцептрона Розенблатта является т. н. проблема XOR: множество нулей и множество единиц данной функции линейно неразделимы. Ввиду линейности конструкции объединение нескольких таких нейронов в сеть не имеет смысла: композиция линейных функций снова будет линейной, и сеть из любого, сколь угодно большого числа линейных перцептронов сможет реализовать только те же самые линейные функции, для которых было достаточно и одного. Для построения нелинейного классификатора вводят нелинейную функцию активации, принимающую на вход линейную комбинацию $\boldsymbol{w}, \boldsymbol{x}\rangle$. Наиболее популярная исторически функция активации — *логистический сигмоид*:

$$f(x) = \frac{1}{1+e^{-x}}.$$

Как и другие функции активации нейронов, это монотонно неубывающая функция, которая при $x \to -\infty$ стремится к нулю, а при $x \to +\infty$ — к единице. Это значит, что при подаче на вход большого по модулю отрицательного числа нейрон не активируется, но активируется при подаче такого же положительного. Более того, данная функция активации равна 0,5 при $x = 0$, что вкупе с использованием *перекрёстной энтропии* в

качестве функции ошибок позволяет получать на выходе не два дискретных значения -1 и +1, а десятичную дробь в диапазоне $(0,5;1)$, означающую вероятность принадлежности $\boldsymbol{x}$ к классу, ассоциированному с конкретным выходом. Средняя перекрёстная энтропия по всем объектам тренирующей выборки имеет вид:

$$E_P(\boldsymbol{w}) = -\frac{1}{n}\sum_{i=1}^{n}[y(\boldsymbol{x}_i)\ln(f(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)) + (1 - y(\boldsymbol{x}_i))\ln(1 - f(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle))].$$

Перейдём от однослойного перцептрона к многослойной полносвязной нейронной сети. Обозначим выход j-того нейрона l-того слоя как y_j^l:

$$y^{j(l)} = f(\sum_{i=0}^{m_0} w_{ji}^l y^{i(l-1)}),$$

где m_0 — размерность l-того слоя, w_{ji}^l — вес синаптической связи, соединяющей j-тый нейрон слоя l с i-тым нейроном слоя $l-1$.

$y^{j(0)} = x^j$, где x^j — j-тый элемент входного вектора.

В данном случае для минимизации значения функции ошибок необходимо подобрать параметры уже не одного нейрона, а всей сети. Для этого используется метод обратного распространения ошибки, являющийся модификацией метода градиентного спуска. В качестве функции ошибок преимущественно используется категориальная перекрёстная энтропия.

2.1.1 Обзор архитектуры свёрточной нейронной сети

Свёрточная нейронная сеть обычно состоит из двух блоков: блока свёрточных слоёв и блока полносвязных слоёв, аналогичных слоям многослойного перцептрона. Типовая архитектура свёрточной нейронной сети приведена на рисунке 2.2.

Основная идея свёрточных слоёв заключается в извлечении карт признаков из двумерных матриц с последующей их передачей в классификатор, в роли которого выступает блок полносвязных слоёв. Карты признаков выгоднее для классификации, чем исходные данные, поскольку они имеют

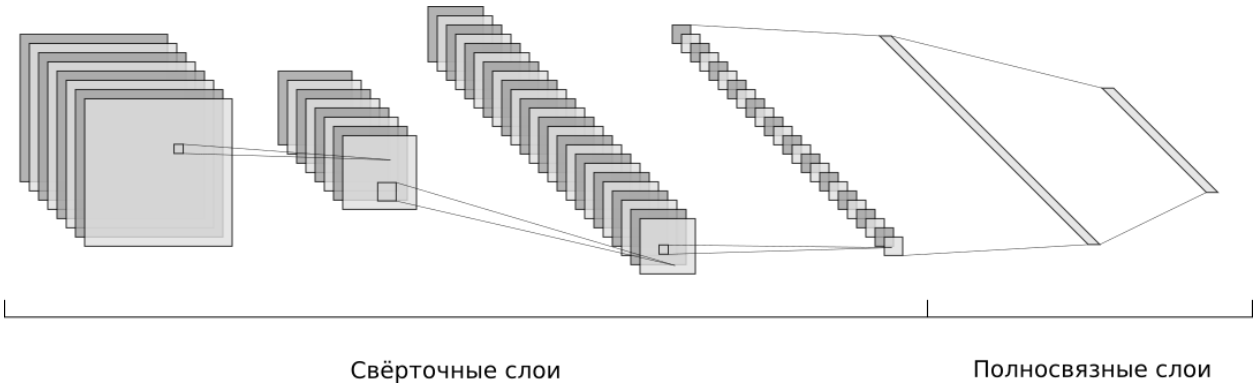


Рисунок 2.2 – Типовая архитектура свёрточной нейронной сети

меньшую размерность, а их содержанием являются наиболее существенные признаки начальных данных. Это объясняет успех свёрточных нейронных сетей в области распознавания образов и классификации изображений.

Каждый свёрточный слой обычно выполняет три операции для получения карт признаков. Первый шаг — фильтрация методом скользящего окна с использованием K ядер, результатом которой является K карт признаков. Каждое ядро применяется к существующей карте признаков, полученной в предыдущем слое. В первом свёрточном слое все ядра применяются к исходному изображению. Затем производится уменьшение размерности карт признаков с помощью операции субдискретизации (пулинга) путём вычисления среднего или максимального значения в каждом регионе карты признаков размера $p \times p$. Финальный шаг — расчёт нелинейной функции карты признаков.

В сравнении с полносвязными слоями свёрточные слои обучаются быстрее ввиду использования метода скользящего окна. Благодаря этому, в обучении нуждаются не все связи между соседними слоями, а только элементы свёрточного ядра, общего для одной конкретной матрицы. Это разделение весов не только существенно уменьшает их количество, но и повышает обобщающие способности сети, позволяя формировать универсальные относительно различных участков изображений свёрточные ядра.

Как и в нейронных сетях других типов, процесс обучения заключается в минимизации функции потерь с использованием оптимизационного алгоритма, обновляющего веса. Пакетный режим стохастического градиентного спуска в качестве оптимизационного алгоритма очень популярен, но AdaDelta или AdaGrad также могут применятся в свёрточных нейронных сетях.

Рассмотрим детальнее некоторые операции, применяемые в свёрточных слоях.

Свёртка. Обозначим результат свёртки в слое l с k-ым ядром, равным матрице весов W^{kl}, как C^{kl}. Тогда

$$C^{kl} = \sum_{m=1}^{K^{l-1}} (W^{kl} * F^{m(l-1)}),$$

где $*$ обозначает операцию свёртки. K^{l-1} — это количество ядер в предыдущем слое, а $F^{m(l-1)}$ — k-ая карта признаков, полученная из предыдущего слоя. Для первого свёрточного слоя $l = 1$, $K^{l-1} = K^0 = 1$, а $F^{1(l-1)} = F^{10} = I$, где I — входное изображение. Размер матрицы фильтрации W^{kl} напрямую определяет размер локальной области (скользящего окна), используемой для вычисления C_{ij}^{kl}.

Размер выходной матрицы также зависит от двух параметров, называемых шагом свёртки и дополнением (англ. padding). Шаг свёртки S задаёт дискрет перемещения скользящего окна, регулирующий степень пересечения локальных областей. Дополнение позволяет дополнить входную карту признаков $F^{m(l-1)}$ или изображение I нулями по краям. Обозначим толщину дополнения P.

$$dim(C^{kl}) = (dim(F^{m(l-1)}) - dim(W^{kl}) + 2 \times P)/S + 1.$$

Например, для сохранения размерности C^{kl} равной размерности входных данных $F^{m(l-1)}$ ($dim(C^{kl}) = dim(F^{m(l-1)})$ необходимо соблюсти условие

$$\begin{cases} S = 1, \\ P = dim(W^{kl}/2). \end{cases}$$

Функция активации. Для внесения нелинейности каждый результат свёртки C^{kl} обрабатывается с помощью функции активации $f^{kl} : \mathbb{R} \to \mathbb{R}$ так же, как это происходит в других типах нейронных сетей. В качестве функции активации в свёрточных нейронных сетях часто используются логистический сигмоид, гиперболический тангенс $f(x) = tanh(x)$ и ReLU:

$$f(x) = max(0, x).$$

Субдискретизация (пулинг). Результатом применения данной операции является уменьшение размерности двумерного массива путём его разбиения на фрагменты размером $p_l \times p_l$ с последующей заменой каждого фрагмента средним или максимальным значением его элементов. Значение шага регулирует пересечение соседних фрагментов, хотя наиболее часто шагу присваивается значение p_l, и пересечения не происходит. Значение p_l обычно выбирается в диапазоне от 2 до 5, в зависимости от размерности входных данных слоя l.

Субдискретизация выполняет две функции: на уровне одного слоя — снижение размерности и путём вычисления «типичного» признака фрагмента или взятия наиболее сильно проявляющегося признака; на уровне сети в целом — агрегирование низкоуровневых признаков в высокоуровневые.

В задачах стегоанализа выбор операции усреднения при осуществлении субдискретизации оправдан слабым характером стеговоздействия, не позволяющим ему достаточно часто принимать максимальное значение на карте признаков.

Запишем выражение для выхода первого свёрточного слоя:

$$F^{k1} = pooling(f^{k1}(\sum_{m=1}^{K^0}(W^{k1} * F^{m(l-1)}) + b^{k1})) = pooling(f^{k1}(W^{k1} * F^{10} + b^{k1}))$$
$$= pooling(f^{k1}(W^{k1} * I + b^{k1})).$$

2.2 Свёрточная нейронная сеть GNCNN

Модель свёрточной нейронной сети для стегоанализа изображений в оттенках серого GNCNN была предложена в [11]. Её основные отличительные особенности: применение фильтра предварительной обработки и использование функции Гаусса в качестве функции активации нейронов свёрточных слоёв.

$$K = \frac{1}{12}\begin{pmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{pmatrix} \tag{2.3}$$

Полосовой высокочастотный фильтр предварительной обработки с импульсной характеристикой (2.3) введён исходя из априорного знания о слабом характере стеговоздействия на стегоконтейнер. Целью предварительной фильтрации является усиление яркости стегосигнала и ослабление яркости оригинального изображения. Веса данного фильтра заранее предопределены и не участвуют в процессе обучения нейронной сети.

Ещё одним преобразованием, облегчающим процесс извлечения признаков, является переход от оригинального изображения к шумоподобному остатку $R = (r_{ij})$, содержащему только предсказания относительно факта стеганографической модификации каждого пикселя:

$$r_{ij} = y_{ij} - P(N(Y, i, j)),$$

где $Y = (y_{ij})$ – стегоконтейнер, а $P(N(Y, i, j))$ – оценка значения пикселя y_{ij}, полученная из окружающих его пикселей $N(Y, i, j)$ [12]. Ввиду наличия сложных зависимостей между соседними пикселями, в случае отсутствия модификации оценка зачастую будет близка к действительному значению пикселя y_{ij}, а значение r_{ij} будет близко к нулю.

$$f(x) = e^{-\frac{x^2}{\sigma^2}} \tag{2.4}$$

Функция Гаусса (2.4) с нулевым математическим ожиданием в качестве функции активации (рис. 2.3) призвана обеспечить формирование в свёрточном слое нейронной сети такого ядра K, что $Y * K = R$. Максимальное значение функции в нуле соответствует нулевой ошибке оценки значения y_{ij} и, следовательно, предположению об отсутствии модификации пикселя стегоконтейнера. Резкий спад по мере удаления от нуля обращает в близкие к нулю значения любой результат свёртки, превышающий порог, определяемый среднеквадратическим отклонением σ, влияющим на ширину кривой.

Архитектура нейронной сети представлена на рис. 2.4 и имеет вид состояний, через которые проходит входное изображение. Между состояниями располагаются порядковые номера слоёв.

На вход нейронной сети подаётся изображение в оттенках серого в разрешении 256×256 пикселей. Слой предварительной обработки 1 с ядром (2.3) не участвует в обучении. За ним следуют пять свёрточных слоёв 2–6, состоящих из 16 каналов. Каждый из них осуществляет операцию свёртки с ядрами, формирующимися по мере обучения сети, а также вычисление функции активации и выполнение субдискретизации по среднему с размером окна 3×3 и шагом 2. Слой 2 использует для свёртки ядра размера 5×5, слои 3–5 — ядра размера 3×3, слой 6 — ядра размера 5×5.

Выход последнего свёрточного слоя представляет собой 256 выделенных признака. Они помещаются в модуль классификации, состоящий из трёх полносвязных слоёв 7–9: первые два имеют по 128 нейронов каждый

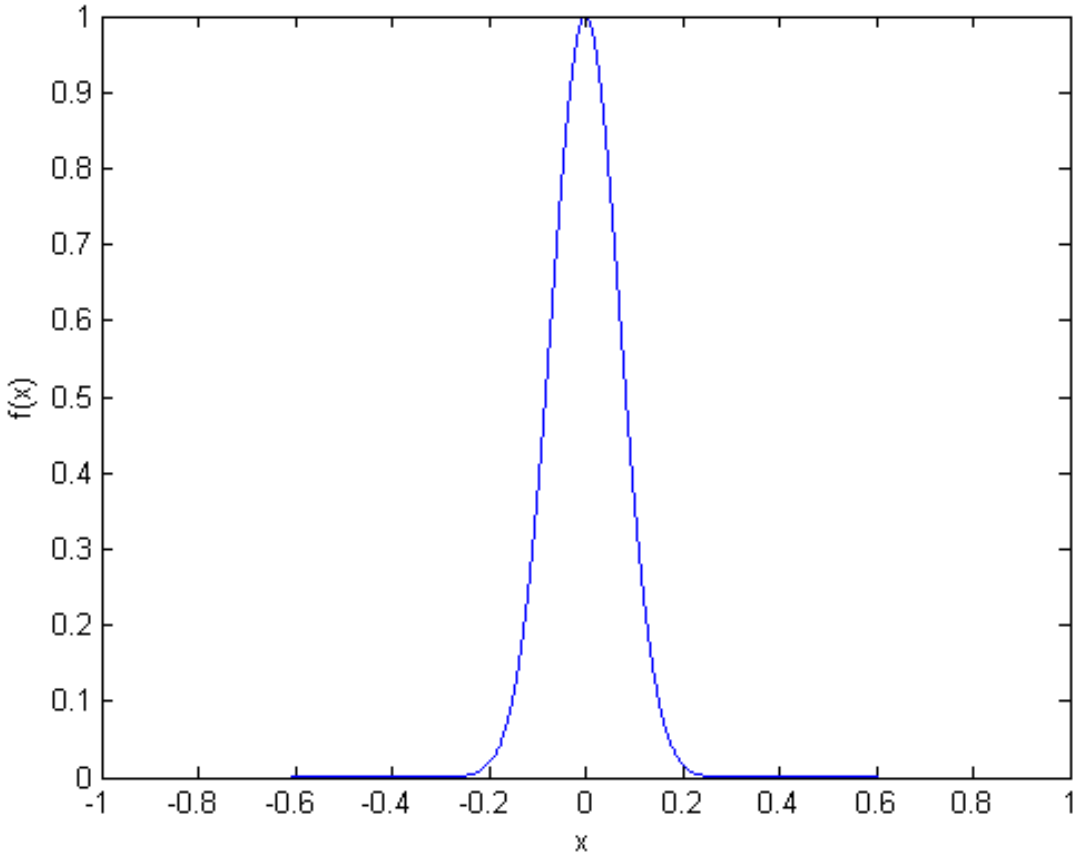


Рисунок 2.3 – Функция активации GNCNN

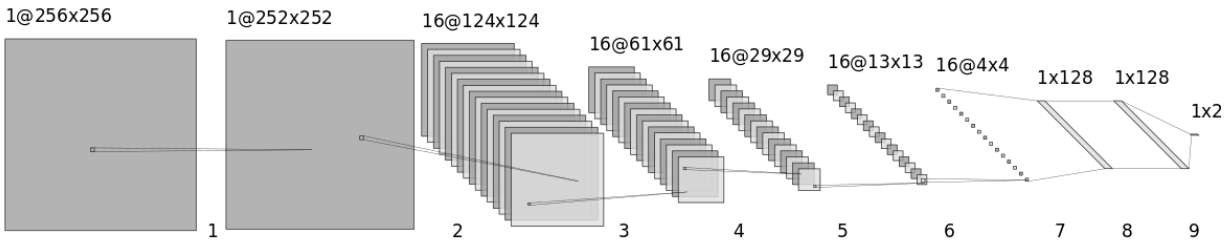


Рисунок 2.4 – Архитектура GNCNN

и функцию активации ReLU, последний – два нейрона и функцию активации softmax. В качестве функции потерь используется перекрёстная энтропия.

2.3 Нейронная сеть с двумя свёрточными слоями

Также в рамках работы была реализована модель свёрточной нейронной сети для стегоанализа изображений в оттенках серого, предложенная в [13]. Её особенностями являются отсутствие процедуры субдискретизации и размер свёрточного слоя, располагающегося непосредственно перед полносвязным:

$$dim(W^{k(L-1)}) = dim(C^{k(L-2)}) - 1.$$

Тогда при $S = 1$ и $P = 0$

$$dim(C^{k(L-1)}) = (dim(C^{k(L-2)}) - (dim(C^{k(L-2)}) - 1)/S + 1 = 2.$$

Таким образом, свёрточный слой $C^{k(L-1)}$ выделяет 4 признака на основе зависимостей во всей карте признаков $dim(F^{k(L-2)}$, а не отдельных фрагментов, что позволяет детектировать алгоритмы, вносящие слабое искажение, неравномерно распределённое по разным частям контейнера, в отличие от блочных стегоалгоритмов.

Для корректности эксперимента размеры слоёв были изменены для работы со входными изображениями в разрешении 256×256 пикселей. Архитектура нейронной сети приведена на 2.5.

На вход нейронной сети подаётся изображение в оттенках серого в разрешении 256×256 пикселей. Первый свёрточный слой имеет размер 3×3. Его цель — произвести высокочастотную фильтрацию подобно тому, как это делает фильтр предварительной обработки GNCNN. Замена константного ядра обучаемым мотивирована отсутствием доказательства оптимальности ядра 2.3. Далее следует свёрточный слой размера 253×253, состоящий из 64 каналов. Оба свёрточных слоя осуществляют операцию свёртки с обучаемым ядром и вычисление функции активации tanh. Результатом их работы

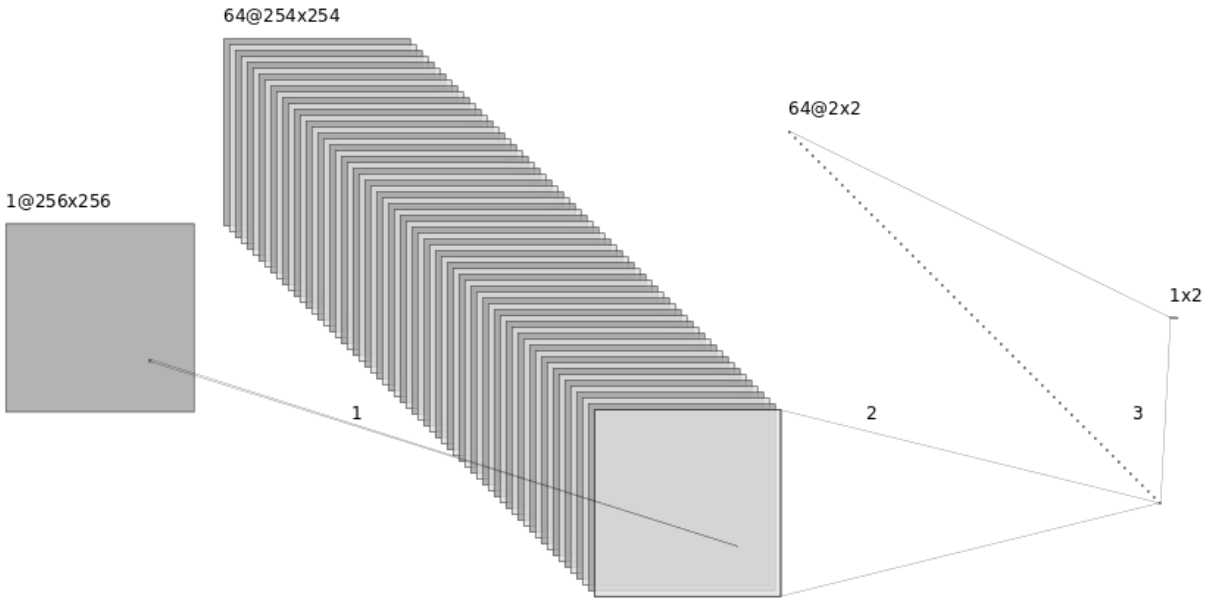


Рисунок 2.5 – Архитектура сети с двумя свёрточными слоями

являются 64 карты признаков размера 2×2, соединённые с двумя нейронами с функцией активации softmax. В качестве функции потерь используется категориальная кросс-энтропия.

2.4 Комбинированная свёрточная нейронная сеть

Рассмотрим архитектуру комбинированной свёрточной сети, совмещающей в себе особенности как GNCNN, так и нейронной сети с двумя свёрточными слоями. Результаты классификации тестовой выборки, приведённые в [13], показывают, что большое количество свёрточных слоёв и уменьшение размерности пространства признаков, создаваемого ими, путём применения субдискретизации не являются необходимыми условиями для построения нейросетевого стегоанализатора, хотя увеличивают вычислительную сложность модели нейронной сети.

Сравнение результатов классификации обеих рассмотренных ранее нейронных сетей, а также сравнение эффекта, достигаемого при использовании различных функций активации совместно с архитектурой GNCNN [14],

также показали, что функция Гаусса в качестве функции активации свёрточных слоёв не даёт никаких преимуществ в сравнении с традиционно используемыми в свёрточных нейронных сетях функциями активации.

Однако, сильной стороной GNCNN является использование фильтра предварительной обработки, увеличивающего скорость сходимости градиентного обучения.

Учитывая вышеперечисленные выводы, для комбинированной свёрточной сети было решено использовать архитектуру нейронной сети с двумя свёрточными слоями, но с добавлением предварительной фильтрации, описанной в [11]. В качестве функции ошибки была выбрана перекрёстная энтропия с L_2-регуляризацией [15]. Конечная архитектура показана на рисунке 2.6.

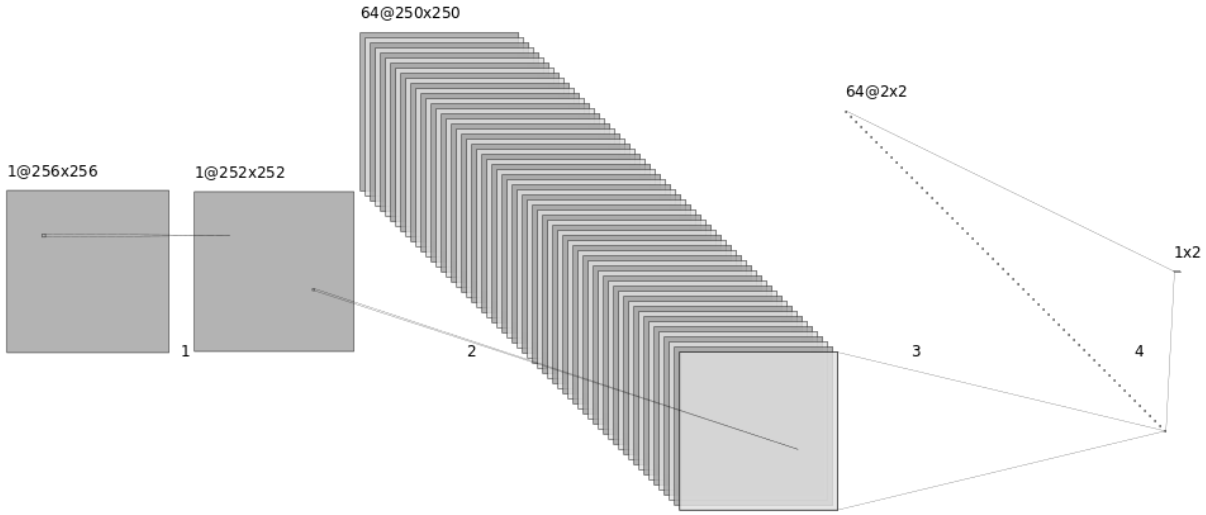


Рисунок 2.6 – Архитектура комбинированной свёрточной сети

3 Описание разработанного программного обеспечения и экспериментальных исследований

В рамках данной работы были реализованы три модели нейронных сетей для стегоанализа изображений в оттенках серого: GNCNN [11], нейронная сеть с двумя свёрточными слоями [13] и комбинированная свёрточная сеть. Впервые были получены свёрточные ИНС для стегоанализа цветных изображений путём модификации вышеперечисленных моделей. Также была произведена оценка способности к различению пустых и заполненных контейнеров для всех шести сетей.

3.1 Структура разработанного программного обеспечения

Программная реализация моделей нейронных сетей выполнена на языке программирования Python в виде шести интерактивных тетрадей IPython Notebook [16]. Работоспособность протестирована для интерпретатора Python версии 3.5.2. При построении нейронных сетей и организации их обучения использованы библиотеки Keras 2.2.4 [17] и TensorFlow 1.13.1 [18]. Для отображения графиков процесса обучения использована библиотека livelossplot 0.3.4 [19].

Фильтр предварительной обработки для GNCNN и комбинированной свёрточной сети реализован в виде отдельного приложения на языке программирования C++. Также для проведения эксперимента был реализован стеганографический метод относительной замены коэффициентов дискретного косинусного преобразования (ДКП) Коха и Жао.

Архитектура модифицированной для работы с цветными изображениями нейронной сети GNCNN приведена на рисунке 3.1, нейронной сети с двумя свёрточными слоями — на рисунке 3.2, комбинированной свёрточной сети — на рисунке 3.3.

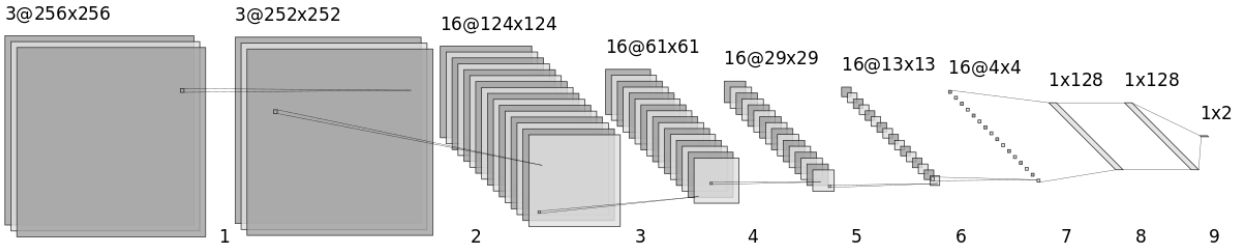


Рисунок 3.1 – Архитектура GNCNN

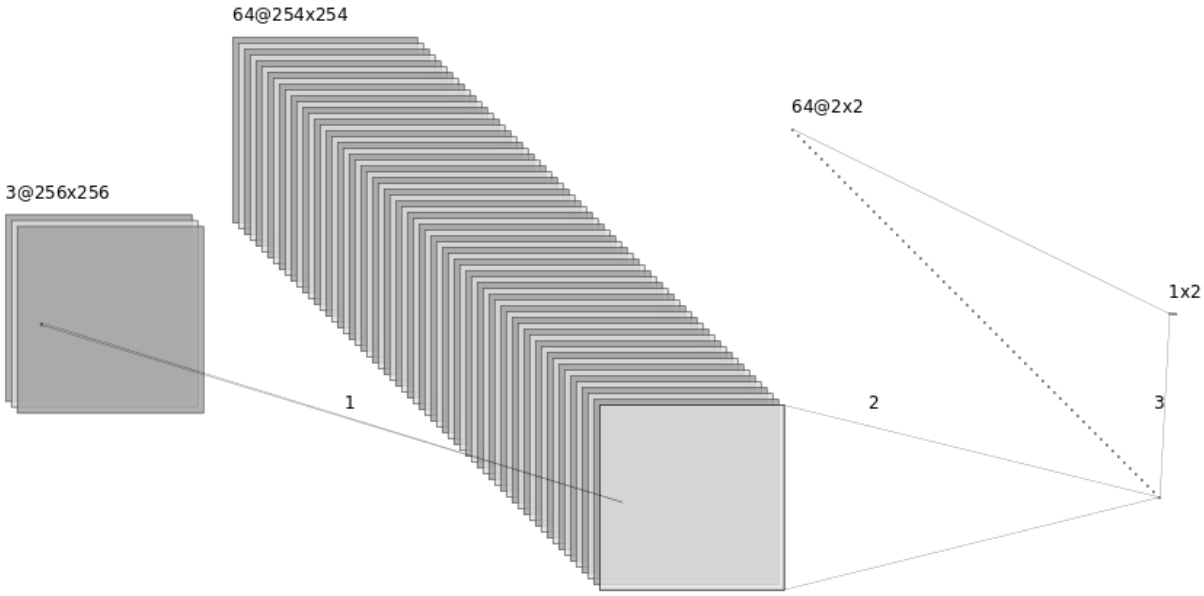


Рисунок 3.2 – Архитектура сети с двумя свёрточными слоями

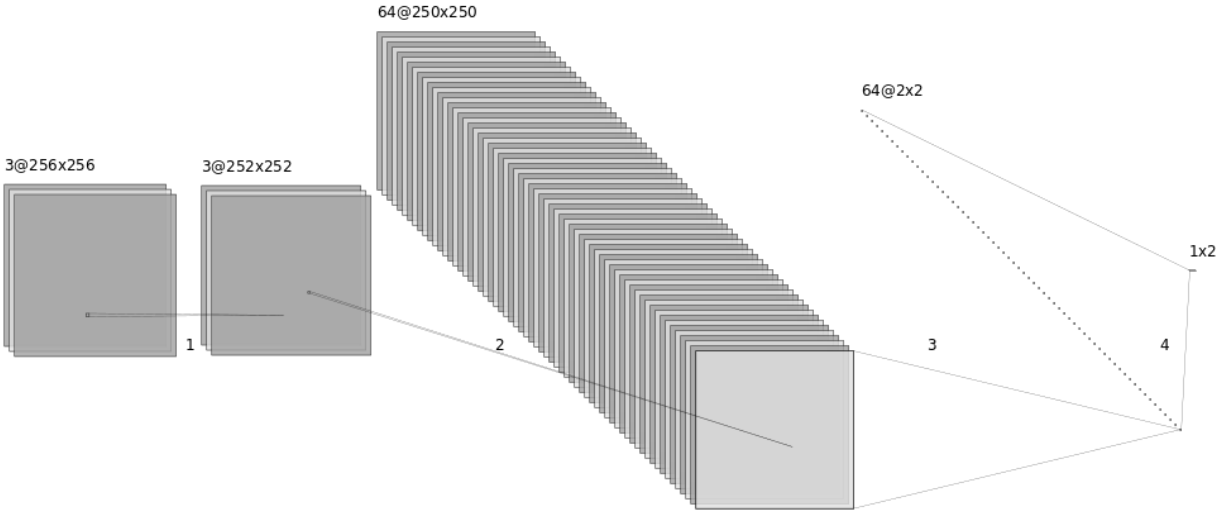


Рисунок 3.3 – Архитектура комбинированной свёрточной сети

Для обучения GNCNN было использовано значение параметра σ функции активации, равное 0,1, и метод обучения Adam [20] со скоростью обу-

чения 0,001, $\beta_1 = 0,9$, $\beta_2 = 0,999$. Для обучения нейронной сети с двумя свёрточными слоями и комбинированной нейронной сети использовался стохастический градиентный спуск со скоростью обучения 0,005.

Блок-схема работы комбинированной свёрточной сети показана на рисунке 3.4.

3.2 Цель, план и результаты эксперимента

Целью проведения эксперимента было определение способности вышеописанных нейронных сетей различать пустые и заполненные различными стеганографическими методами контейнеры. Для заполнения использовалось следующее программное обеспечение:

– программная реализация метода создания цифровых водяных знаков на основе гетероассоциативных сжимающих преобразований (ГСП) [21],

– собственная реализация метода относительной замены коэффициентов дискретного косинусного преобразования (ДКП) Коха и Жао [22, 23],

– симулятор стеговстраивания с использованием алгоритма WOW [24].

Выбор алгоритмов стеговстраивания обусловлен следующими предпосылками:

– метод создания цифровых водяных знаков на основе гетероассоциативных сжимающих преобразований (ГСП) является блочным, вносит малые искажения в контейнер и ранее не использовался в задачах стегоанализа с применением свёрточных нейронных сетей,

– метод относительной замены коэффициентов дискретного косинусного преобразования (ДКП) Коха и Жао вносит существенные искажения и может служить индикатором общей работоспособности алгоритмов стегоанализа,

– алгоритм WOW [24] не является блочным и признан среди современных стеганографических алгоритмов как один из наиболее скрытных.

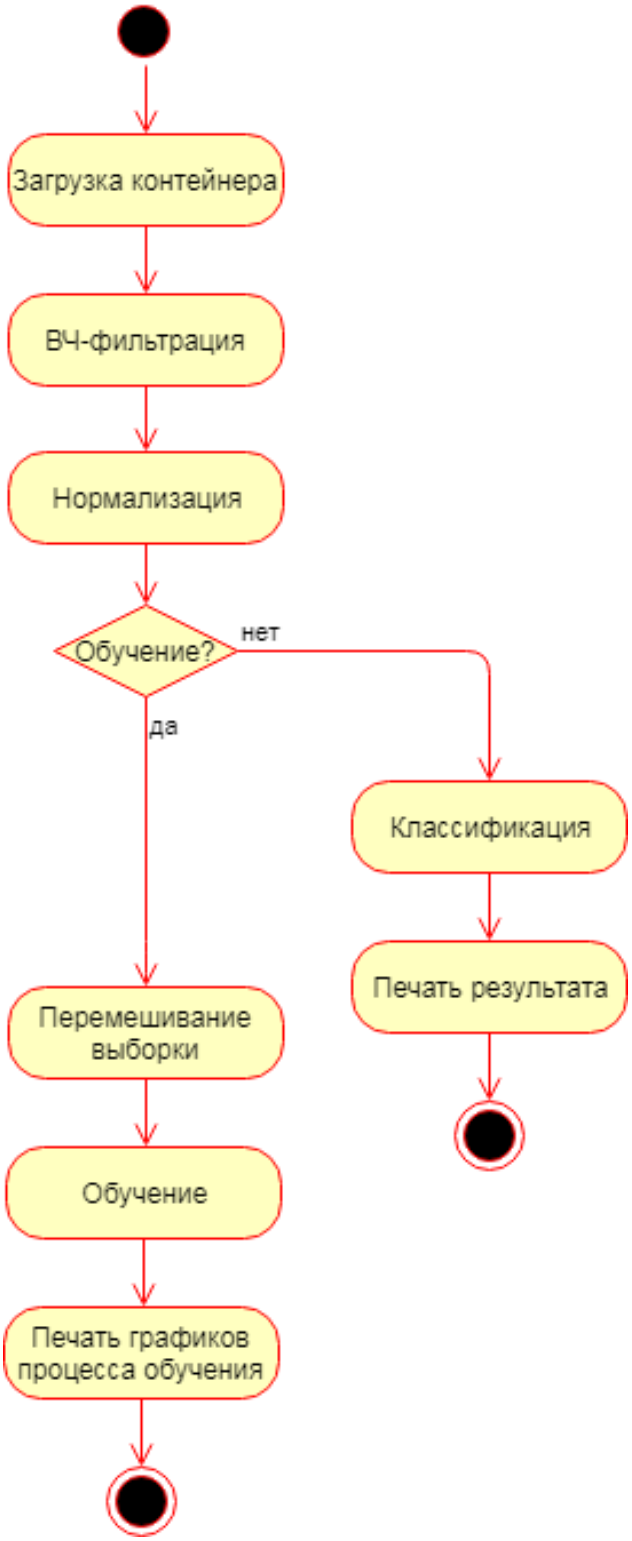


Рисунок 3.4 – Блок-схема работы комбинированной свёрточной сети

Для проведения эксперимента использовалась база данных изображений PPG-LIRMM-COLOR [25], состоящая из 10 000 изображений в разрешении 512×512 пикселей. База данных подверглась конвертации в формат изображений в оттенках серого с помощью утилиты ppmtopgm [26]. Затем из каждого изображения были вырезаны два фрагмента размером 256×256 пикселей: один из них использовался для обучения нейронной сети сжатия в составе реализации метода создания цифровых водяных знаков на основе ГСП, второй – для осуществления встраивания обученной сетью. В дальнейшем последний фрагмент использовался для стеговстраивания с применением собственной реализации метода Коха и Жао, а также симулятора встраивания с использованием алгоритма WOW.

Таким образом, для обучения нейронных сетей были получены три выборки из 10 тыс. пустых и 10 тыс. заполненных одним из перечисленных стегоалгоритмов контейнеров в оттенках серого. Такие же выборки были получены для цветных изображений из оригинальной базы.

Ввиду отсутствия поддержки симулятором встраивания с использованием алгоритма WOW цветных изображений, пустые контейнеры предварительно были разделены по каналам цветности на три изображения, затем была произведена симуляция встраивания в изображения, соответствующие синем каналу, и наконец, разделённые изображения были вновь собраны в цветные заполненные контейнеры.

Кроме того, были сформированы аналогичные выборки из 1 тыс. пустых и 1 тыс. заполненных контейнеров для оценки влияния объёма выборки на точность классификации.

Обучающая подвыборка составила 90 % полученной выборки, валидационная — 10 % во всех случаях.

В рамках эксперимента было произведено сравнение реализаций методов стеговстраивания по средней среднеквадратической ошибке и среднему проценту восстановленных посредством стегоизвлечения данных (таблица 1

для контейнеров в оттенках серого и таблица 2 для цветных), а также обучение рассмотренных нейронных сетей с целью оценки их способности к классификации стегоконтейнеров валидационной подвыборки (таблица 3 для контейнеров в оттенках серого и таблица 4 для цветных).

Для каждого стегоалгоритма указан параметр, характеризующий мощность стеговоздействия. Для метода на основе гетероассоциативных сжимающих преобразований им является амплитуда встраивания A_m, для алгоритма Коха и Жао – разность между коэффициентами ДКП p, кодирующая различие между логическими нулём и единицей встраиваемой битовой последовательности, для алгоритма WOW – количество встроенных битов, делённое на количество пикселей в изображении, α.

Таблица 1 – Сравнение методов встраивания в контейнеры в оттенках серого

Стегоалгоритм	MSE	Процент восстановленных данных
На основе ГСП ($A_m = 0,016$)	0,1334	97,23 %
Алгоритм Коха и Жао ($p = 1$)	0,5775	99,78 %
WOW ($\alpha = 0,4$)	0,0389	–

Таблица 2 – Сравнение методов встраивания в цветные контейнеры

Стегоалгоритм	MSE	Процент восстановленных данных
На основе ГСП ($A_m = 0,016$)	0.0475	98,79 %
Алгоритм Коха и Жао ($p = 1$)	0.1995	99,76 %
WOW ($\alpha = 0,4$)	0.0129	–

Процент успешно извлечённых после стеговстраивания данных для алгоритма WOW не указан ввиду использования симулятора, не производящего сокрытия информации, а лишь изменяющего пиксели соответственно алгоритму встраивания.

Таблица 3 – Точность классификации контейнеров в оттенках серого

Стегоалгоритм	GNCNN (2 тыс.)	GNCNN (20 тыс.)	2-хслойная НС (2 тыс.)	2-хслойная НС (20 тыс.)	Комб. НС (2 тыс.)	Комб. НС (20 тыс.)
На основе ГСП ($A_m = 0,016$)	94,1 %	96,8 %	48 %	51,7 %	96 %	95,1 %
Алгоритм Коха и Жао ($p = 1$)	50,7 %	93,1 %	100 %	99,8 %	100 %	99 %
WOW ($\alpha = 0,4$)	50 %	49,6 %	85,5 %	96,3 %	92,5 %	95 %

Таблица 4 – Точность классификации цветных контейнеров

Стегоалгоритм	GNCNN (2 тыс.)	GNCNN (20 тыс.)	2-хслойная НС (2 тыс.)	2-хслойная НС (20 тыс.)	Комб. НС (2 тыс.)	Комб. НС (20 тыс.)
На основе ГСП ($A_m = 0,016$)	53,7 %	99,5 %	50 %	50 %	82,5 %	100 %
Алгоритм Коха и Жао ($p = 1$)	49,3 %	49,1 %	99,5 %	100 %	98 %	98,1 %
WOW ($\alpha = 0,4$)	50 %	49,4 %	54,5 %	96,7 %	90 %	91,1 %

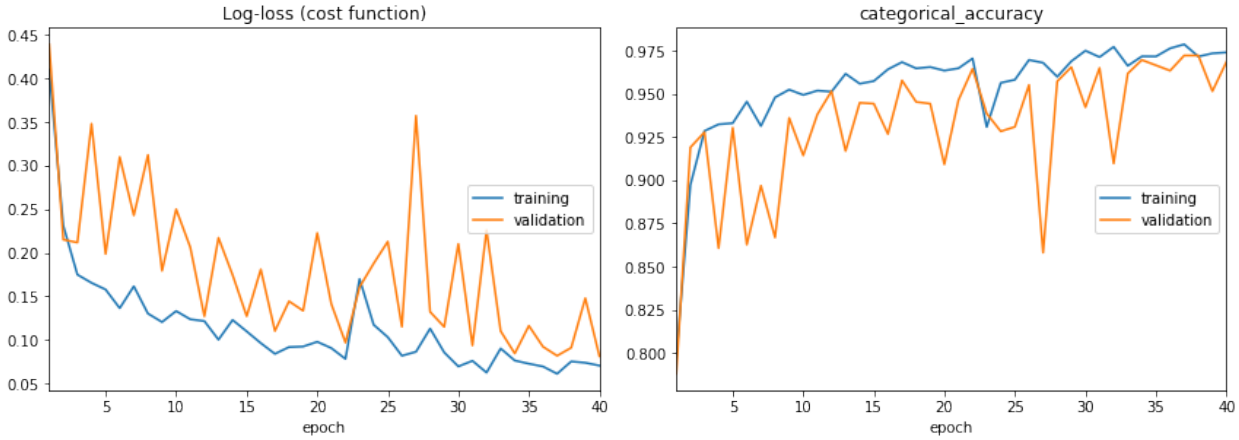
Из таблицы 3 видно, что все три нейронные сети имеют практически одинаковую способность к различению пустых контейнеров и контейнеров, заполненных с применением метода Коха и Жао, на выборке из 20 тысяч изображений в оттенках серого.

Однако, нейронная сеть с двумя свёрточными слоями оказалась неспособна успешно детектировать контейнеры, заполненные с помощью метода на основе ГСП, а GNCNN — заполненные с применением симулятора встраивания WOW, в то время как комбинированная свёрточная сеть одинаково хорошо справилась с обеими задачами.

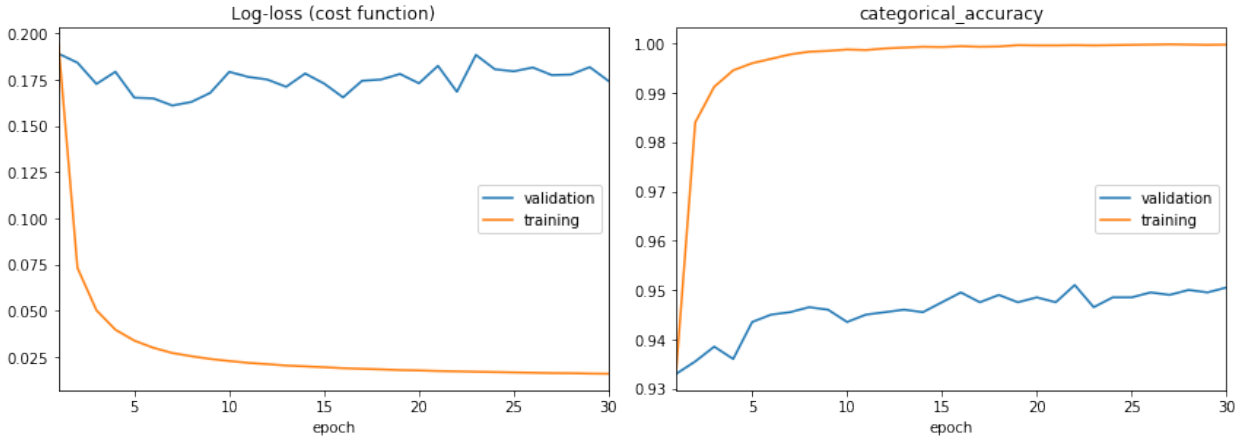
Объём выборки не возымел существенного влияния на точность классификации во всех случаях, кроме распознавания контейнеров, заполненных методом Коха и Жао, нейронной сетью GNCNN.

Из таблицы 4 видно, что GNCNN распознаёт цветные контейнеры хуже, чем контейнеры в оттенках серого: для алгоритма на основе ГСП проявляется влияние объёма выборки, метод Коха и Жао перестаёт детектироваться вовсе. Различающая способность нейронной сети с двумя свёрточными слоями также начинает зависеть от объёма выборки для алгоритма WOW. Комбинированная свёрточная сеть ухудшает свои показатели по сравнению с результатами, полученными на выборке в оттенках серого, но точность классификации остаётся приемлемой.

Рисунки 3.5–3.10 иллюстрируют процесс обучения нейронных сетей. На них изображены графики зависимости функции ошибок и точности классификации от количества эпох. Графики для тренировочной и валидационной подвыборок обозначены в легенде как «training» и «validation» соответственно. Помимо прочего, графики иллюстрируют более гладкий характер кривых для комбинированной свёрточной сети, обусловленный использованием L_2-регуляризации.

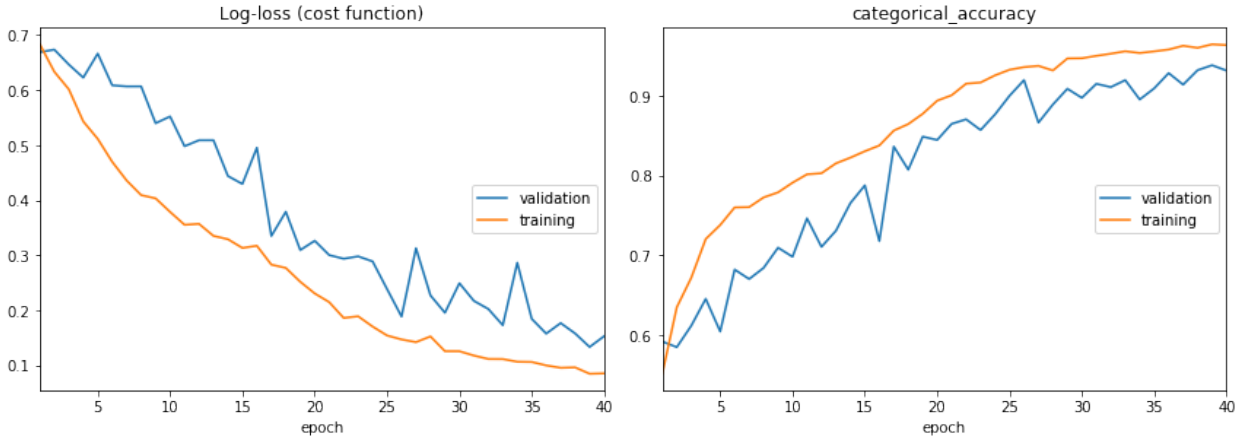


(а) GNCNN

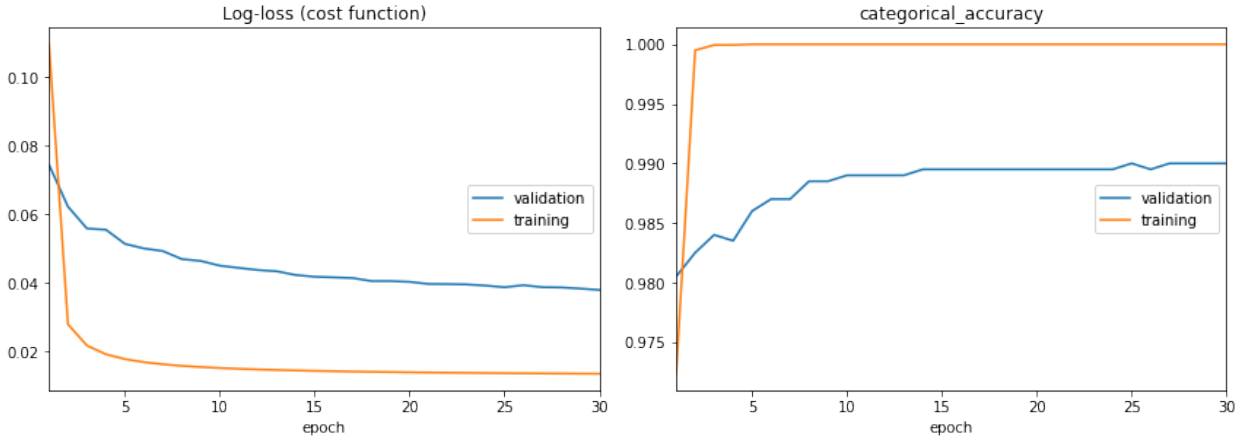


(б) Комбинированная СНС

Рисунок 3.5 – Графики процесса обучения для ГСП в оттенках серого

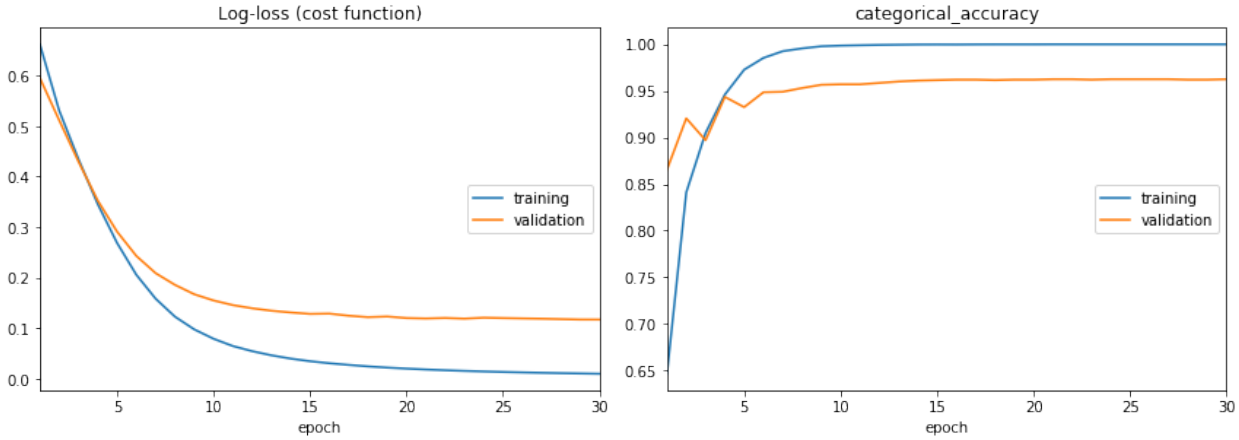


(а) GNCNN

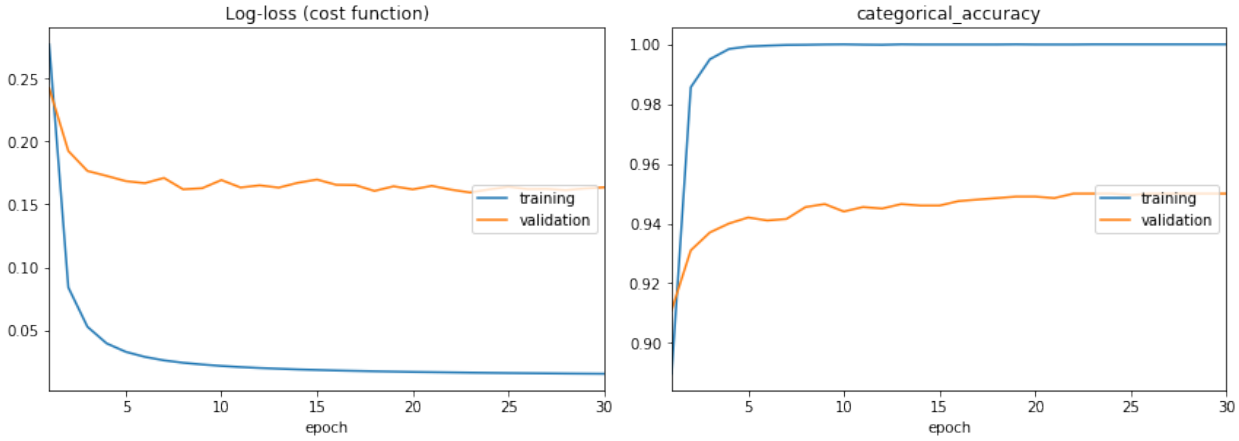


(б) Комбинированная CHC

Рисунок 3.6 – Графики процесса обучения для Коха и Жао в оттенках серого

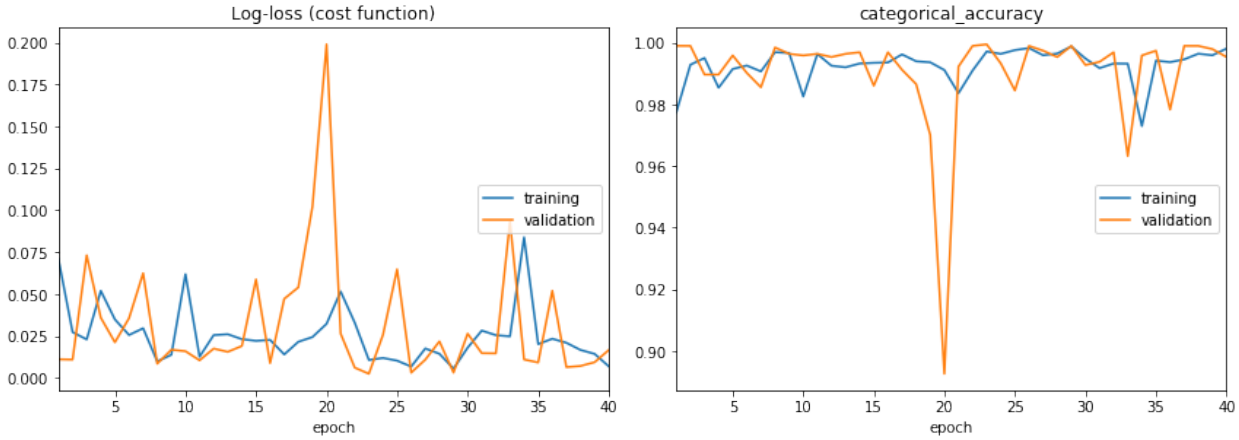


(а) 2-хслойная НС

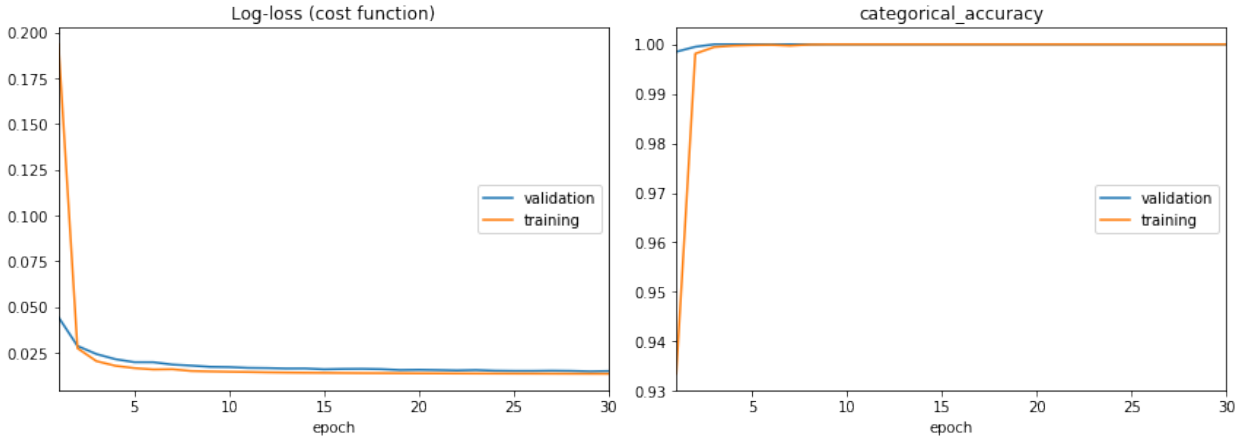


(б) Комбинированная СНС

Рисунок 3.7 – Графики процесса обучения для WOW в оттенках серого

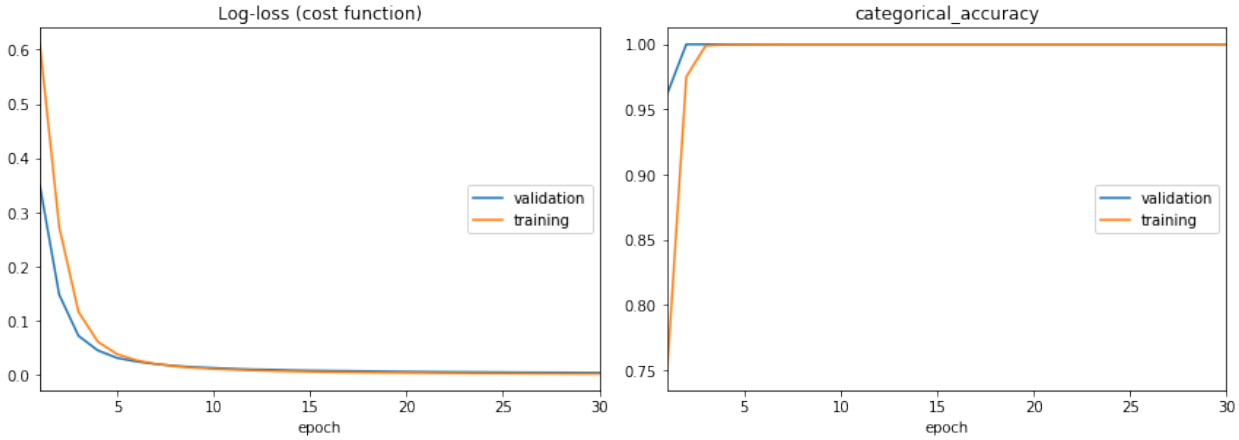


(а) GNCNN

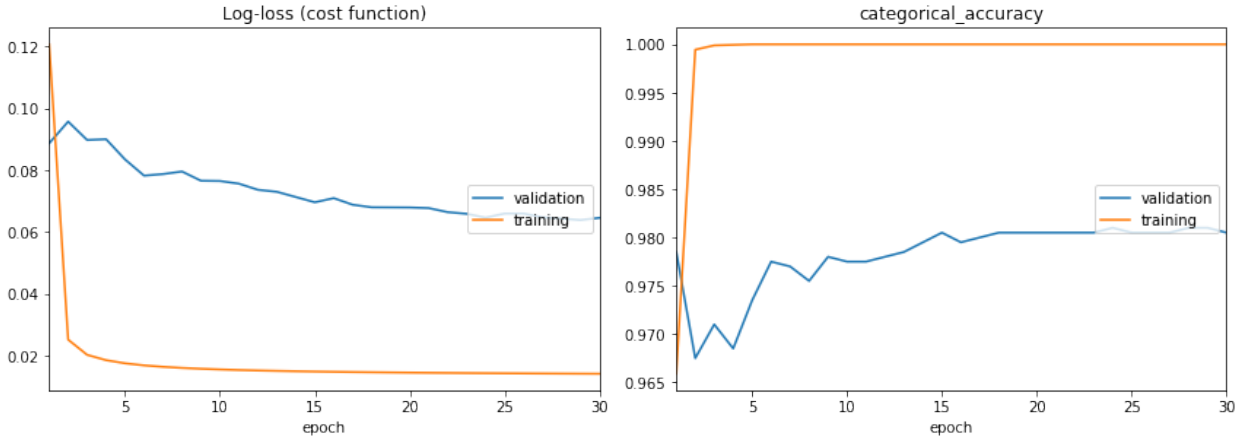


(б) Комбинированная CHC

Рисунок 3.8 – Графики процесса обучения для цветного ГСП

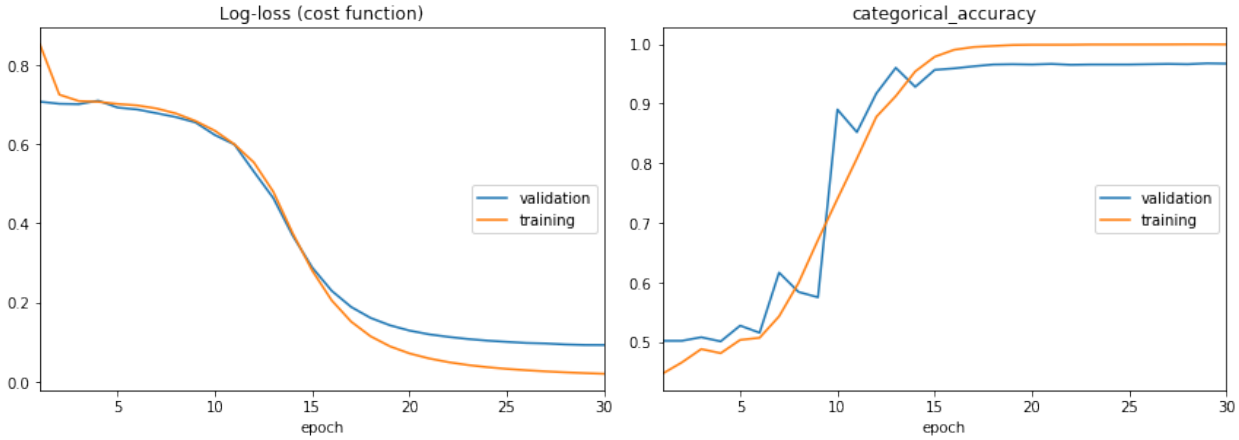


(а) 2-хслойная НС

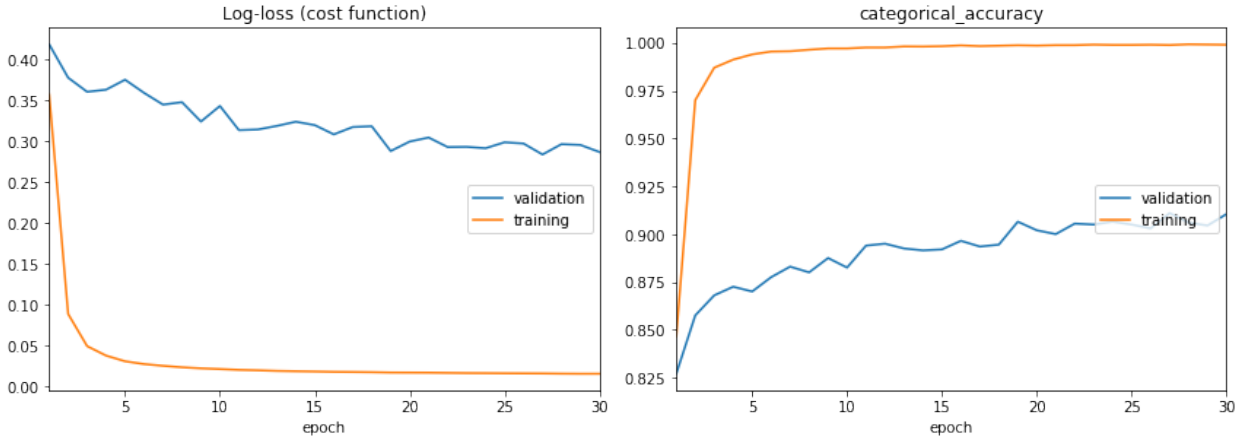


(б) Комбинированная СНС

Рисунок 3.9 – Графики процесса обучения для цветного Коха и Жао



(а) 2-хслойная НС



(б) Комбинированная СНС

Рисунок 3.10 – Графики процесса обучения для цветного WOW

Заключение

В рамках данной работы был разработан и реализован алгоритм стегоанализа цветных изображений с использованием свёрточной нейронной сети. Кроме того, были реализованы известные нейросетевые алгоритмы стегоанализа изображений в оттенках серого GNCNN и нейронная сеть с двумя свёрточными слоями, впоследствии модифицированные для работы с цветными изображениями.

Исходя из результатов сравнения вышеупомянутых стегоаналитических алгоритмов, можно сделать предположение о том, что ввиду использования в GNCNN скользящих окон малого размера данная модель нейронной сети лучше подходит для анализа контейнеров, заполненных при помощи блочных алгоритмов, а нейронная сеть с двумя свёрточными слоями ввиду большого размера окна свёртки лучше подходит для стегоанализа алгоритмов, неравномерно распределяющих стегосигнал в пространственной области контейнера.

На примере комбинированной свёрточной сети показана практическая эффективность процедуры предварительной высокочастотной фильтрации. Направлением дальнейших исследований может быть поиск критерия оптимальности фильтра предварительной обработки.

Дополнительного исследования также требует факт более устойчивого детектирования алгоритмов, производящих встраивание в три канала цветного изображения, чем алгоритмов, производящих встраивание только в один канал, нейронной сетью GNCNN.

Список используемых источников

1. Стеганография, цифровые водяные знаки и стеганоанализ: Монография / А. В. Аграновский, А. В. Балакин, В. Г. Грибунин, С. А. Сапожников. — М. : Вузовская книга, 2009. — 220 с.

2. Fridrich J., Goljan M., Du R. Lossless Data Embedding—New Paradigm in Digital Watermarking // EURASIP Journal on Advances in Signal Processing. — 2002. — Vol. 2002, no. 2.

3. Fridrich J. Steganography in Digital Media: Principles, Algorithms, and Applications. — 1st edition. — New York : Cambridge University Press, 2009. — 437 p.

4. Johnson N. F., Jajodia S. Steganalysis of Images Created Using Current Steganography Software // Information Hiding. — Springer Berlin Heidelberg, 1998. — P. 273–289.

5. Fridrich J., Goljan M., Du R. Detecting LSB steganography in color, and gray-scale images // IEEE Multimedia. — 2001. — Vol. 8, no. 4. — P. 22–28.

6. Detecting Original Image Using Histogram, DFT and SVM / T. H. Manjula Devi, H.S. Manjunatha Reddy, K. B. Raja et al. // International Journal of Recent Trends in Engineering. — 2009. — Vol. 1, no. 1. — P. 367–371.

7. Rishidas S., Gayathri Krishnan L., Sujith Kumar T. P. A Comparative Study of Steganalysis using Support Vector Machines on Different Image Formats // International Journal of Engineering Research and Technology. — 2015. — Vol. 4, no. 3.

8. Rumelhart D. E., Hinton G. E., Williams R. J. Learning Internal Representations by Error Propagation // Computational models of cognition

and perception. — Cambridge, MA : MIT Press, 1986. — Vol. 1. — P. 319–362.

9. Rosenblatt F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain // Psychological Review. — 1958. — Vol. 65, no. 6. — P. 386–408.

10. Rosenblatt F. Principles of Neurodynamics. — New York : Spartan, 1962.

11. Deep learning for steganalysis via convolutional neural networks / Y. Qian, J. Dong, W. Wang, T. Tan // Media Watermarking, Security, and Forensics 2015 / Ed. by A. M. Alattar, N. D. Memon, C. D. Heitzenrater. — SPIE, 2015.

12. Steganalysis of Content-Adaptive Steganography in Spatial Domain / J. Fridrich, J. Kodovský, V. Holub, M. Goljan // Information Hiding. — Springer Berlin Heidelberg, 2011. — P. 102–117.

13. Steganalysis via a Convolutional Neural Network using Large Convolution Filters [Electronic resource online] / J.-F. Couchot, R. Couturier, C. Guyeux, M. Salomon // CoRR. — 2016. — Access mode: http://arxiv.org/abs/1605.07946.

14. Feature learning for steganalysis using convolutional neural networks / Y. Qian, J. Dong, W. Wang, T. Tan // Multimedia Tools and Applications. — 2017. — Vol. 77, no. 15. — P. 19633–19657.

15. Тихонов А. Н. О некорректных задачах линейной алгебры и устойчивом методе их решения // ДАН СССР. — 1965. — Vol. 163. — P. 591–594.

16. Jupyter and the future of IPython — IPython [Electronic resource]. — Access mode: https://ipython.org.

17. Home - Keras Documentation [Electronic resource]. — Access mode: https://keras.io.

18. TensorFlow [Electronic resource]. — Access mode: https://www.tensorflow.org.

19. stared/livelossplot: Live training loss plot in Jupyter Notebook for Keras, PyTorch and others [Electronic resource]. — Access mode: https://github.com/stared/livelossplot.

20. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization // CoRR. — 2015. — Access mode: https://arxiv.org/abs/1412.6980.

21. Сирота А. А., Дрюченко М. А., Митрофанова Е. Ю. Метод создания цифровых водяных знаков на основе гетероассоциативных сжимающих преобразований изображений и его реализация с использованием искусственных нейронных сетей // Компьютерная оптика. — 2018. — № 3. — С. 483–494.

22. Zhao J., Koch E. Embedding Robust Labels Into Images For Copyright Protection // Proceedings of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies. — 1995. — P. 242–251.

23. Koch E., Zhao J. Towards Robust and Hidden Image Copyright Labeling // Proc. of 1995 IEEE Workshop on Nonlinear Signal and Image Processing. — 1995. — P. 452–455.

24. Holub V., Fridrich J. Designing steganographic distortion using directional filters // 2012 IEEE International Workshop on Information Forensics and Security (WIFS). — IEEE, 2012.

25. PPG-LIRMM-COLOR base [Electronic resource]. — Access mode: http://www.lirmm.fr/~chaumont/PPG-LIRMM-COLOR.html.

26. Ppmtopgm User Manual [Electronic resource]. — Access mode: http:// netpbm.sourceforge.net/doc/ppmtopgm.html.